# Stochastic optimal control theory

Bert Kappen
SNN, Radboud University
Nijmegen the Netherlands

August, 2012

# SNN Machine Learning and Brain modeling

- Develop mathematical methods for the brain and intelligent behavior
  - Bayesian methods for learning and data analysis
  - Control theory
  - Applications

- Approach
  - methods from statistical physics, statistics, computer science, mathematics
  - insights from neuroscience

PhD position available on Neural Networks for stochastic optimal control theory

# www.snn.ru.nl

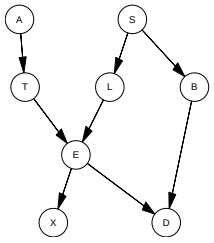
**Bayesian methods**
- Graphical models
- Approximate inference

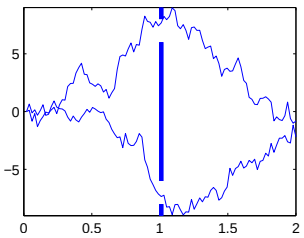
**Stochastic optimal control theory**
- Controlled diffusions
- Learning
- Robotics and multi-agents

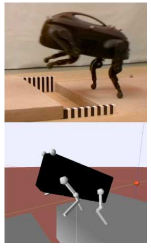
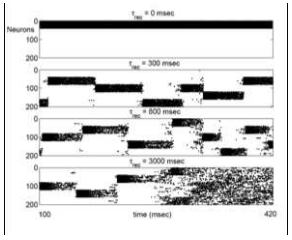
**Bio-informatics**
- Genome-wide association studies
- Neuro-imaging genetics

**Neuroscience**
- Neural networks
- Brain computer interface
- Connectivity measures

**Spin-off** smart-research.nl
- Wine portal
- Petrophysical expert system (Shell)
- Medical diagnostic expert system
- Missing person identification (NFI)





# Introduction


**Optimal control theory:** Optimize sum of a path cost and end cost. Result is optimal control sequence and optimal trajectory.

**Input:** Cost function. **Output:** Optimal trajectory and controls.

# Introduction

Control problems are delayed reward problems:

- Motor control: devise a sequece of motor commands to reach a goal
- finance: devise a sequence of buy/sell commands to maximize profit
- Learning, exploration exploitation

# Types of optimal control problems

Finite horizon (fixed horizon time)

- Dynamics and environment may depend explicitly on time.
- Optimal control depends explicitly on time.

# Types of optimal control problems

Finite horizon (moving horizon)

- Dynamics and environment are static.
- Optimal control is time independent.

Infinite horizon

- discounted reward, Reinforcement learning
- total reward, absorbing states
- average reward

Other issues:

- discrete vs. continuous state
- discrete vs. continuous time
- observable vs. partial observable

# Overview

**Lecture 1: Optimal control theory, discrete time**

- Introduction of delayed reward problem in discrete time;
- Dynamic programming solution and deterministic Bellman equations;
- Extension to noisy case
- Examples

# Overview

**Lecture 2: Optimal control theory, continuous time**

- Solution in continuous time and states;
- Example: Mass on a spring
- Pontryagin maximum principle; Notion of an optimal (particle) trajectory
- Again Mass on a spring
- Stochastic differential equations
- Kolmogorov and Fokker-Plack equations

# Overview

**Lecture 3: Stochastic optimal control theory**

- Hamilton-Jacobi-Bellman equation (continuous state and time)
- LQ control, Ricatti equation;
- Example of LQ control
- Portfolio selection
- Path integral control

# Overview

**Lecture 4: KL control theory**

- Example: Delayed choice
- Importance sampling
- How to control a device?
- KL control theory - Relation KL control and path integral control - Multi agent system - Stationary KL control (- Variational approximation, $n$ joint arm)
(- Coordination of continuous agents using MF and BP)
(- Risk sensitive control)
(- Inference and control)

# Material

- H.J. Kappen. Optimal control theory and the linear Bellman Equation. In *Inference and Learning in Dynamical Models (Cambridge University Press 2010)*, edited by David Barber, Taylan Cemgil and Sylvia Chiappa
  `http://www.snn.ru.nl/~bertk/control/timeseriesbook.pdf`
- Dimitri Bertsekas, Dynamic programming and optimal control
- website

# Lecture 1: Optimal control theory: discrete time

# Discrete time control

Consider the control of a discrete time deterministic dynamical system:

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

$x_t$ describes the *state* and $u_t$ specifies the *control* or *action* at time $t$.

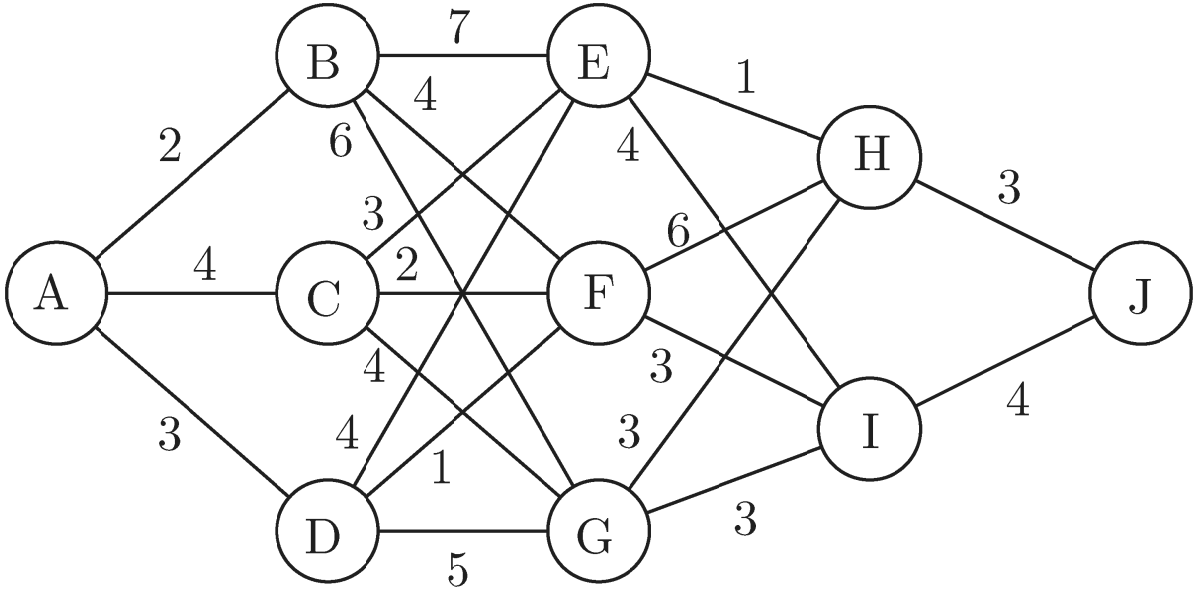
Given $x_{t=0} = x_0$ and $u_{0:T-1} = u_0, u_1, \ldots, u_T - 1$, we can compute $x_{1:T}$.

Define a cost for each sequence of controls:

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

The problem of optimal control is to find the sequence $u_{0:T-1}$ that minimizes $C(x_0, u_{0:T-1})$.

# Dynamic programming




Find the minimal cost path from A to J.

$$
\begin{aligned}
C(J) &= 0, C(H) = 3, C(I) = 4 \\
C(F) &= \min(6 + C(H), 3 + C(I))
\end{aligned}
$$

# Discrete time control

The optimal control problem can be solved by dynamic programming. Introduce the *optimal cost-to-go*:

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

which solves the optimal control problem from an intermediate time $t$ until the fixed end time $T$, for all intermediate states $x_t$.

Then,

$$\begin{aligned} J(T, x) &= \phi(x) \\ J(0, x) &= \min_{u_{0:T-1}} C(x, u_{0:T-1}) \end{aligned}$$

# Discrete time control

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

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

This is called the *Bellman Equation*.

Computes $u$ as a function of $x,t$ for all intermediate $t$ and all $x$.

# Discrete time control

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

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

NB: the backward computation requires $u^*_t(x)$ for all $x$.

# Stochastic case

$$x_{t+1} = x_t + f(t, x_t, u_t, w_t) \qquad t = 0, \ldots, T-1$$

At time $t$, $w_t$ is a random value drawn from a probability distribution $p(w)$.

For instance,

$$\begin{aligned} x_{t+1} &= x_t + w_t, \qquad x_0 = 0 \\ w_t &= \pm 1, \qquad p(w_t = 1) = p(w_t = -1) = 1/2 \\ x_t &= \sum_{s=0}^{t-1} w_s \end{aligned}$$

Thus, $x_t$ random variable and so is the cost

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

# Stochastic case

$$
\begin{aligned}
C(x_0) &= \left\langle \phi(x_T) + \sum_{t=0}^{T-1} R(t, x_t, u_t, \xi_t) \right\rangle \\
&= \sum_{w_{0:T-1}} \sum_{\xi_{0:T-1}} p(w_{0:T-1}) p(\xi_{0:T-1}) \left( \phi(x_T) + \sum_{t=0}^{T-1} R(t, x_t, u_t, \xi_t) \right)
\end{aligned}
$$

with $\xi_t, x_t, w_t$ random. Closed loop control: find *functions* $u_t(x_t)$ that minimizes the remaining expected cost when in state $x$ at time $t$. $\pi = \{u_0(\cdot), \ldots, u_{T-1}(\cdot)\}$ is called a policy.

$$
\begin{aligned}
x_{t+1} &= x_t + f(t, x_t, u_t(x_t), w_t) \\
C_\pi(x_0) &= \left\langle \phi(x_T) + \sum_{t=0}^{T-1} R(t, x_t, u_t(x_t), \xi_t) \right\rangle
\end{aligned}
$$

$\pi^* = \mathrm{argmin}_\pi C_\pi(x_0)$ is optimal policy.

# Stochastic Bellman Equation

$$
\begin{aligned}
J(t, x_t) &= \min_{u_t} \langle R(t, x_t, u_t, \xi_t) + J(t+1, x_t + f(t, x_t, u_t, w_t)) \rangle \\
J(T, x) &= \phi(x)
\end{aligned}
$$

$u_t$ is optimized for each $x_t$ separately. $\pi = \{u_0, \ldots, u_{T-1}\}$ is optimal a policy.

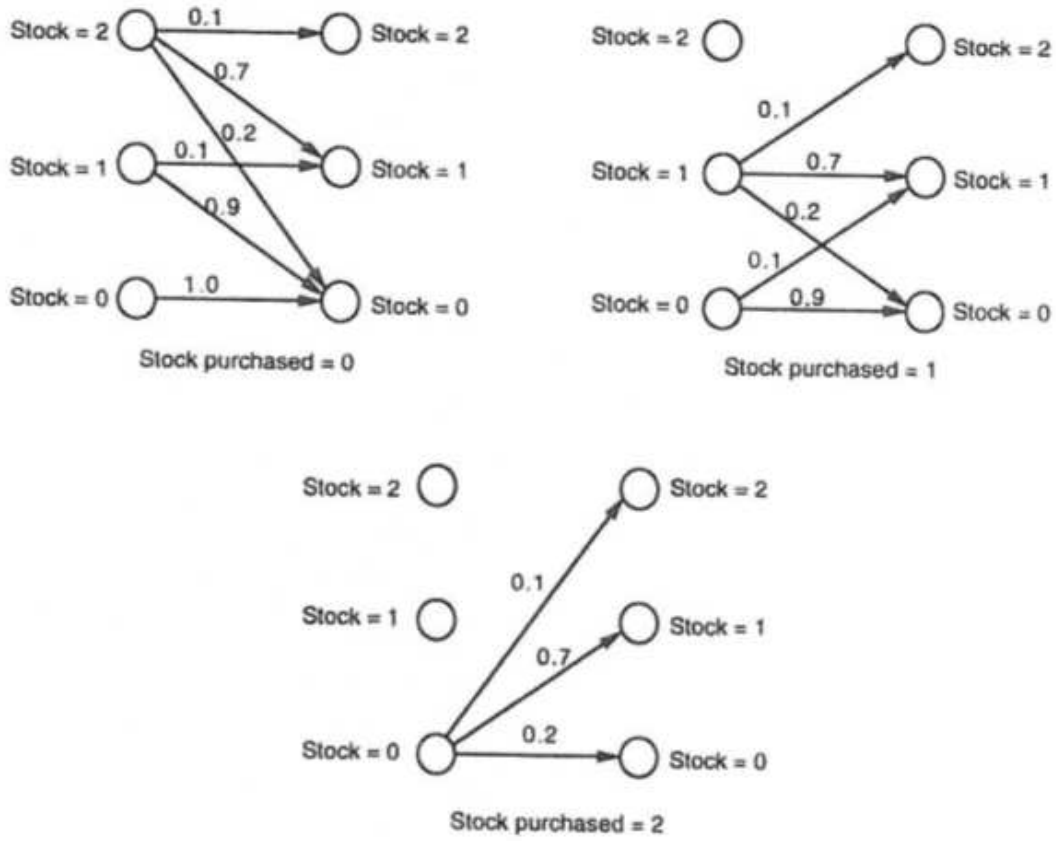
# Inventory problem

- $x_t = 0, 1, 2$ stock available at the beginning of period $t$.
- $u_t$ stock ordered at the beginning of period $t$. Maximum storage is 2: $u_t \leq 2 - x_t$.
- $w_t = 0, 1, 2$ demand during period $t$ with $p(w = 0, 1, 2) = (0.1, 0.7, 0.2)$; excess demand is lost.
- $u_t$ is the cost of purchasing $u_t$ units. $(x_t + u_t - w_t)^2$ is cost of stock at end of period $t$.

$$
\begin{aligned}
x_{t+1} &= \max(0, x_t + u_t - w_t) \\
C(x_0, u_{0:T-1}) &= \left\langle \sum_{t=0}^{t=2} u_t + (x_t + u_t - w_t)^2 \right\rangle
\end{aligned}
$$

Planning horizon $T = 3$.

# Inventory problem



# Apply Bellman Equation

$$
\begin{aligned}
J_t(x_t) &= \min_{u_t} \langle R(x_t, u_t, w_t) + J_{t+1}(f(x_t, u_t, w_t)) \rangle \\
R(x, u, w) &= u + (x + u - w)^2 \\
f(x, u, w) &= \max(0, x + u - w)
\end{aligned}
$$

Start with $J_3(x_3) = 0, \forall x_3$.

# Dynamic programming in action

Assume we are at stage $t = 2$ and the stock is $x_2$. The cost-to-go is what we order $u_2$ and how much we have left at the end of period $t = 2$.

$$\begin{aligned} J_2(x_2) &= \min_{0 \le u_2 \le 2 - x_2} u_2 + \left\langle (x_2 + u_2 - w_2)^2 \right\rangle \\ &= \min_{0 \le u_2 \le 2 - x_2} \big( u_2 + 0.1 * (x_2 + u_2)^2 + 0.7 * (x_2 + u_2 - 1)^2 \\ &+ 0.2 * (x_2 + u_2 - 2)^2 \big) \\ J_2(0) &= \min_{0 \le u_2 \le 2} \big( u_2 + 0.1 * u_2^2 + 0.7 * (u_2 - 1)^2 + 0.2 * (u_2 - 2)^2 \big) \end{aligned}$$

$$\begin{aligned} u_2 = 0 \quad &: \quad rhs = 0 + 0.7 * 1 + 0.2 * 4 = 1.5 \\ u_2 = 1 \quad &: \quad rhs = 1 + 0.1 * 1 + 0.2 * 1 = 1.3 \\ u_2 = 2 \quad &: \quad rhs = 2 + 0.1 * 4 + 0.7 * 1 = 3.1 \end{aligned}$$

Thus, $u_2(x_2 = 0) = 1$ and $J_2(x_2 = 0) = 1.3$

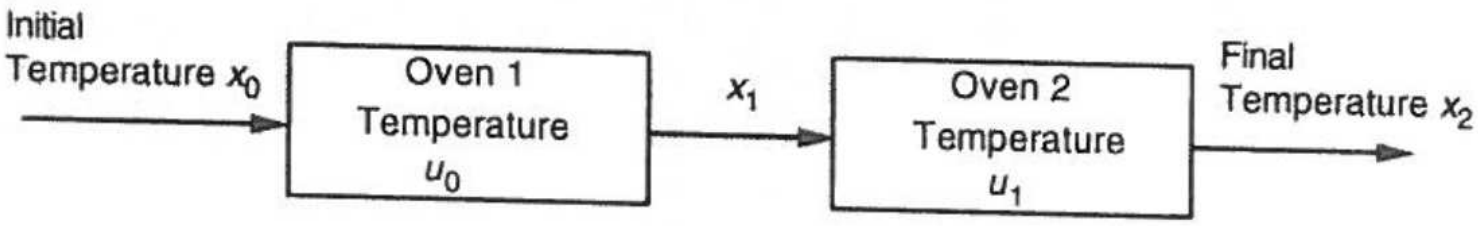
# Inventory problem

The computation can be repeated for $x_2 = 1$ and $x_2 = 2$, completing stage 2 and subsequently for stage 1 and stage 0.

| Stock | Stage 0 Cost-to-go | Stage 0 Optimal stock to purchase | Stage 1 Cost-to-go | Stage 1 Optimal stock to purchase | Stage 2 Cost-to-go | Stage 2 Optimal stock to purchase |
|---|---|---|---|---|---|---|
| 0 | 3.67 | 1 | 2.5 | 1 | 1.3 | 1 |
| 1 | 2.67 | 0 | 1.2 | 0 | 0.3 | 0 |
| 2 | 2.608 | 0 | 1.68 | 0 | 1.1 | 0 |

# Exercise: Two ovens

A certain material is passed through a sequence of two ovens. Aim is to reach pre-specified final product temperature $x^*$ with minimal oven energy.




$x_{0,1,2}$ are the product temperatures initially, after pasing through oven 1 and after passing through oven 2. $u_{0,1}$ are the oven temperatures. The dynamics is

$$\begin{aligned} x_{t+1} &= (1-a)x_t + au_t \qquad t = 0, 1 \\ C &= r(x_2 - x^*)^2 + u_0^2 + u_1^2 \end{aligned}$$

- Find the optimal control solution $u_0, u_1$.
- Show that adding mean zero noise to the dynamics ($x_{t+1} = (1-a)x_t + au_t + w_t$ with $\langle w_t \rangle = 0$), does not change the optimal control solution.

# Example: Two ovens

End cost-to-go is $J(2, x_2) = r(x_2 - x^*)^2$.

$$
\begin{aligned}
J(1, x_1) &= \min_{u_1} \left( u_1^2 + J(2, x_2) \right) = \min_{u_1} \left( u_1^2 + r((1-a)x_1 + au_1 - x^*)^2 \right) \\
u_1 &= \mu_1(x_1) = \frac{ra(x^* - (1-a)x_1)}{1 + ra^2} \\
J(1, x_1) &= \frac{r((1-a)x_1 - x^*)^2}{1 + ra^2} \\
J(0, x_0) &= \min_{u_0} \left( u_0^2 + J(1, x_1) \right) = \min_{u_0} \left( u_0^2 + \frac{r((1-a)x_1 - x^*)^2}{1 + ra^2} \right) \\
&= \min_{u_0} \left( u_0^2 + \frac{r((1-a)((1-a)x_0 + au_0) - x^*)^2}{1 + ra^2} \right) \\
u_0 &= \mu_0(x_0) = \frac{r(1-a)a(x^* - (1-a)^2 x_0)}{1 + ra^2(1 + (1-a)^2)} \\
J(0, x_0) &= \frac{r((1-a)^2 x_0 - x^*)^2}{1 + ra^2(1 + (1-a)^2)}
\end{aligned}
$$

# Comments

- **Linear Quadratic Control:** Solution can be obtained in closed form because problem is linear quadratic.
- **Certainty equivalence:** Optimal control solution is unaffected by noise:

$$\begin{aligned} x_{t+1} &= (1-a)x_t + au_t + w_t \qquad t = 0, 1 \\ C &= r(x_2 - x^*)^2 + u_0^2 + u_1^2 \end{aligned}$$

with $\langle w_t \rangle = 0$.Then

$$\begin{aligned} J(1, x_1) &= \min_{u_1} \left( u_1^2 + \left\langle r((1-a)x_1 + au_1 + w_1 - x^*)^2 \right\rangle \right) \\ &= \min_{u_1} \left( u_1^2 + r((1-a)x_1 + au_1 - x^*)^2 + r \left\langle w_1 \right\rangle^2 \right) \end{aligned}$$

# Lecture 2: Optimal control theory, continuous time

# Continuous limit

Replace $t+1$ by $t+dt$ with $dt \to 0$.

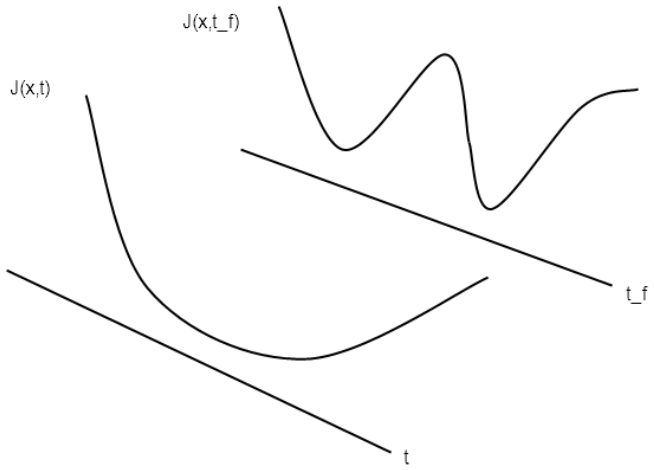
$$\begin{aligned} x_{t+dt} &= x_t + f(x_t, u_t, t)dt \\ C(x_0, u_{0\to T}) &= \phi(x_T) + \int_0^T d\tau R(\tau, x(\tau), u(\tau)) \end{aligned}$$

Assume $J(x,t)$ is smooth.

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

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

# Continuous limit



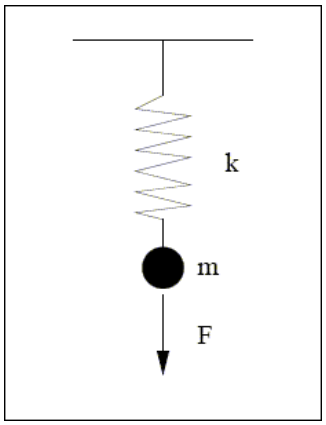

$$-\partial_t J(t,x) \quad = \quad \min_u \left( R(t,x,u) + f(x,u,t)\partial_x J(x,t) \right)$$

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

This is called the *Hamilton-Jacobi-Bellman Equation*.

Computes the *anticipated potential* $J(t,x)$ from the future potential $\phi(x)$.

# Example: Mass on a spring




The spring force $F_z = -z$ towards the rest position and control force $F_u = u$.

Newton's Law

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

with $m = 1$.

Control problem: Given initial position and velocity $z(0) = \dot{z}(0) = 0$ at time $t = 0$, find the control path $-1 < u(0 \to T) < 1$ such that $z(T)$ is maximal.

# Example: Mass on a spring

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

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -x_1 + u \end{aligned}$$

The end cost is $\phi(x) = -x_1$; path cost $R(x, u, t) = 0$.

The HJB takes the form:

$$\begin{aligned} -\partial_t J &= \min_u \left( x_2 \frac{\partial J}{\partial x_1} - x_1 \frac{\partial J}{\partial x_2} + \frac{\partial J}{\partial x_2} u \right) \\ &= x_2 \frac{\partial J}{\partial x_1} - x_1 \frac{\partial J}{\partial x_2} - \left| \frac{\partial J}{\partial x_2} \right|, \quad u = -\text{sign} \left( \frac{\partial J}{\partial x_2} \right) \end{aligned}$$

# Example: Mass on a spring

We try $J(t,x) = \psi_1(t)x_1 + \psi_2(t)x_2 + \alpha(t)$. The HJBE reduces to the ordinary differential equations

$$\begin{aligned} \dot{\psi}_1 &= \psi_2 \\ \dot{\psi}_2 &= -\psi_1 \\ \dot{\alpha} &= -|\psi_2| \end{aligned}$$

These equations must be solved for all $t$, with final boundary conditions $\psi_1(T) = -1, \psi_2(T) = 0$ and $\alpha(T) = 0$.

Note, that the optimal control only requires $\partial_x J(x,t)$, which in this case is $\psi(t)$ and thus we do not need to solve $\alpha$. The solution for $\psi$ is

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

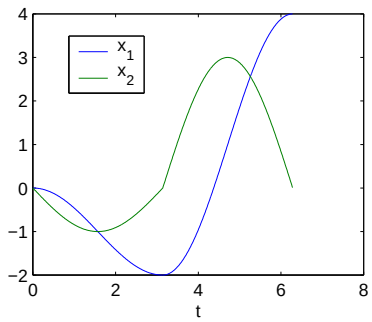
# Example: Mass on a spring

The optimal control is

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

As an example consider $T = 2\pi$. Then, the optimal control is

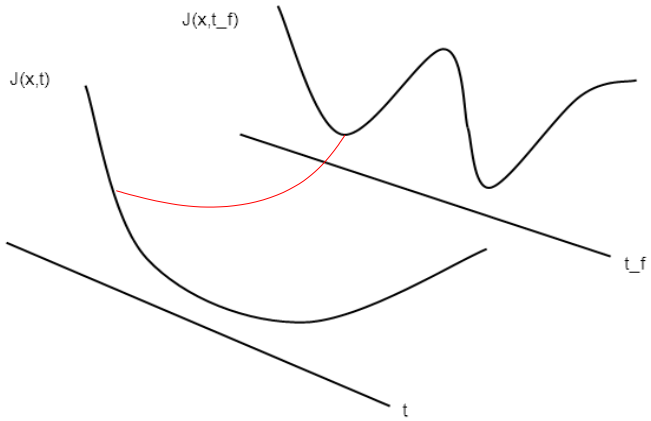
$$u = -1, \quad 0 < t < \pi$$
$$u = 1, \quad \pi < t < 2\pi$$



# Pontryagin minimum principle

The HJB equation is a PDE with boundary condition at future time. The PDE is solved using discretization of space and time.

The solution is an optimal cost-to-go for all $x$ and $t$. From this we compute the optimal trajectory and optimal control.




An alternative approach is a variational approach that directly finds the optimal trajectory and optimal control.

# Pontryagin minimum principle

We can write the optimal control problem as a constrained optimization problem with independent variables $u(0 \to T)$ and $x(0 \to T)$

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

subject to the constraint

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

and boundary condition $x(0) = x_0$.

Introduce the Lagrange multiplier function $\lambda(t)$:

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

# Derivation PMP

The solution is found by extremizing $\mathcal{C}$. This gives a necessary but not sufficient condition for a solution.

If we vary the action wrt to the trajectory $x$, the control $u$ and the Lagrange multiplier $\lambda$, we get:

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

For instance, $H_x = \frac{\partial H(t,x(t),u(t),\lambda(t))}{\partial x(t)}$.

We can solve $H_u(t, x, u, \lambda) = 0$ for $u$ and denote the solution as

$$u^*(t, x, \lambda)$$

Assumes $H$ convex in $u$.

The remaining equations are

$$\begin{aligned} \dot{x} &= H_\lambda(t, x, u^*(t, x, \lambda), \lambda) \\ \dot{\lambda} &= -H_x(t, x, u^*(t, x, \lambda), \lambda) \end{aligned}$$

with boundary conditions

$$x(0) = x_0 \qquad \lambda(T) = -\phi_x(x(T))$$

Mixed boundary value problem.

# Again mass on a spring

Problem

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

Hamiltonian

$$\begin{aligned} H(t,x,u,\lambda) &= -R(t,x,u) + \lambda^T f(t,x,u) = \lambda_1 x_2 + \lambda_2(-x_1 + u) \\ H^*(t,x,\lambda) &= \lambda_1 x_2 - \lambda_2 x_1 - |\lambda_2| \qquad u^* = -\mathsf{sign}(\lambda_2) \end{aligned}$$

The Hamilton equations

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

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

# Example

Consider the control problem:

$$\begin{aligned} dx &= udt \\ C &= \frac{\alpha}{2}x(T)^2 + \int_{t_0}^{T} dt \frac{1}{2}u(t)^2 \end{aligned}$$

with initial condition $x(t_0)$.

Solve the control problem using the PMP formalism.

# Solution

The PMP recipe is

1. Construct the Hamiltonian

$$H(t,x,u,\lambda) = -R(t,x,u) + \lambda f(t,u,x) = -\frac{1}{2}u^2 + \lambda u$$

2. Construct the optimized Hamiltonian

$$H^*(t,x,\lambda) = H(t,x,u^*,\lambda) = \frac{1}{2}\lambda^2 \qquad u^* = \lambda$$

3. Solve the Hamilton equations of motion

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H^*}{\partial \lambda} = \lambda \\ \frac{d\lambda}{dt} &= -\frac{\partial H^*}{\partial x} = 0 \end{aligned}$$

with boundary conditions $x(t_0)$ and $\lambda(t = T) = -\alpha x(T)$[1]. The solution for $\lambda$ is constant $\lambda(t) = \lambda = -\alpha x(T)$. The solution for $x(t)$ is

$$x(t) \quad = \quad x(t_0) + \lambda(t - t_0)$$

Combining these two results, we get $\lambda = -\alpha x(T) = -\alpha(x(t_0) + \lambda(T - t_0))$, or

$$\lambda = \frac{-\alpha x(t_0)}{1 + \alpha(T - t_0)}$$

Since $u^* = \lambda$, this is the optimal control law.

[1] Note, that $\phi(x) = \frac{\alpha}{2}x^2$ so that $\phi_x = \alpha x$.

# Brownian bridge

Due to certainty equivalence, this is also the optimal control law for

$$\begin{aligned} dx &= udt + d\xi \\ C &= \left\langle \frac{\alpha}{2} x(T)^2 + \int_{t_0}^{T} dt \frac{1}{2} u(t)^2 \right\rangle \end{aligned}$$

For $\alpha \to \infty$ the process is known as a Brownian bridge.

The control law and dynamics becomes

$$\begin{aligned} dx &= udt + d\xi \\ u &= \frac{-x(t_0)}{T - t_0} \end{aligned}$$

$x(T) \to 0$ w.p. 1.

# Relation to classical mechanics

The equations look like classical mechanics

$$\begin{aligned} \dot{x} &= H_\lambda(t, x, u^*(t, x, \lambda), \lambda) \qquad x(0) = x_0 \\ \dot{\lambda} &= -H_x(t, x, u^*(t, x, \lambda), \lambda) \qquad \lambda(T) = -\phi_x(x(T)) \end{aligned}$$

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

$$\begin{aligned} \dot{H} &= H_t + H_u \dot{u} + H_x \dot{x} + H_\lambda \dot{\lambda} = H_t \\ H(t, x, u, \lambda) &= -R(t, x, u) + \lambda f(t, u, x) \end{aligned}$$

So, for problems where $R, f$ do not explicitly depend on time, $H$ is a constant of the motion.

# Example

Consider the control problem:

$$\begin{aligned} dx &= udt \\ C &= \int_{t_0}^{T} dt \frac{1}{2} u(t)^2 + V(x(t)) \end{aligned}$$

with initial condition $x(t_0)$.

1. $H(x, u, \lambda) = -\frac{1}{2}u^2 - V(x) + \lambda u$
2. $u^* = \lambda$, $H^*(x, \lambda) = \frac{1}{2}\lambda^2 - V(x)$
3. 

$$\dot{x} = \frac{\partial H^*}{\partial \lambda} = \lambda \qquad \dot{\lambda} = -\frac{\partial H^*}{\partial \lambda} = \frac{\partial V(x)}{\partial x}$$

Control cost $V$ play role of *minus* potential energy.
Control solution has constant *difference* of kinetic energy and state cost

# 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.

Optimal control in continuous space and time contains many complications related to the existence, uniqueness and smoothness of the solution, particular in the absence of noise. In the presence of noise many of these intricacies disappear.

HJB generalizes to the stochastic case, PMP does not (at least not easy).

# Stochastic differential equations

Consider the random walk on the line:

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

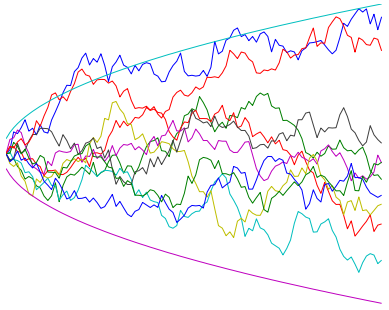
with $x_0 = 0$. We can compute

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

Since $x_t$ is a sum of random variables, $x_t$ becomes Gaussian distributed with

$$\langle x_t \rangle = \sum_{i=1}^{t} \langle \xi_i \rangle = 0$$

$$\langle x_t^2 \rangle = \sum_{i,j=1}^{t} \langle \xi_i \xi_j \rangle = \sum_{i=1}^{t} \langle \xi_i^2 \rangle + \sum_{i,j=1, j\neq i}^{t} \langle \xi_i \xi_j \rangle = t$$



Note, that the fluctuations $\propto \sqrt{t}$.

# Stochastic differential equations

In the continuous time limit we define

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

with $d\xi$ an infinitesimal mean zero Gaussian variable with $\left\langle d\xi^2 \right\rangle = \nu dt$.

Then

$$\frac{d}{dt}\left\langle x \right\rangle \quad = \quad \lim_{dt\to 0}\left\langle \frac{x_{t+dt}-x_t}{dt} \right\rangle = \lim_{dt\to 0}\left\langle \frac{d\xi}{dt} \right\rangle = 0$$

$$\frac{d}{dt}\left\langle x^2 \right\rangle \quad = \quad \lim_{dt\to 0}\left\langle \frac{x^2_{t+dt}-x^2_t}{dt} \right\rangle = \lim_{dt\to 0}\left\langle \frac{(x_t+d\xi)^2-x^2_t}{dt} \right\rangle = \lim_{dt\to 0}\left\langle \frac{d\xi^2}{dt} \right\rangle = \nu$$

So for initial state $x_0$, $\left\langle x \right\rangle(t) = x_0$ and $\left\langle x^2 \right\rangle(t) = \nu t$ which fully specifies the Gaussian distribution:

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

Consider the stochastic differential equation

$$x(t+dt) = x(t) + f(x(t),t)dt + \xi(t)$$

$\xi$ is a Wiener process with $\langle \xi \rangle = 0, \langle \xi^2 \rangle = \nu dt$.

The probability to find the particle at $y$ at time $t+dt$ given that it was at $x$ at time $t$ is given by

$$p(y,t+dt|x,t) \quad = \quad \langle \delta(y-x-f(x,t)dt-\xi) \rangle_\xi$$

where $\langle \rangle_\xi$ is expectation wrt the Wiener process.

# Kolmogorov backward equation

Define $\psi(x,t) = p(z,T|x,t)$ the probability to reach a future state $z$ at time $T$, given that it is currently at $x,t$. Clearly,

$$\begin{aligned}
\psi(x,t) &= p(z,T|x,t) = \int dy p(z,T|y,t+dt)p(y,t+dt|x,t) \\
&= \int dy \psi(y,t+dt) \left\langle \delta(y-x-f(x,t)dt-\xi) \right\rangle_\xi \\
&= \left\langle \psi(x+f(x,t)dt+\xi,t+dt) \right\rangle_\xi \\
&= \psi(x,t) + dt\partial_t\psi(x,t) + \left\langle f(x,t)dt+\xi \right\rangle_\xi \nabla\psi(x,t) \\
&+ \frac{1}{2}\left\langle (f(x,t)dt+\xi)^2 \right\rangle_\xi \nabla^2\psi(x,t)
\end{aligned}$$

Thus,

$$-\partial_t\psi(x,t) = f(x,t)\nabla\psi(x,t) + \frac{1}{2}\nu\nabla^2\psi(x,t) \qquad \psi(x,T) = \delta(z-x)$$

This equation is known as the Kolmogorov backwards equation.

# Fokker Plank (forward) equation

We can similarly derive a forward equation for the quantity $\rho(x,t) = p(x,t|x_0,0)$.

$$\begin{aligned}
\rho(y,t+dt) &= \int dx p(y,t+dt|x,t)\rho(x,t) \\
&= \int dx \left\langle \delta(y-x-f(x,t)dt-\xi)\right\rangle_\xi \rho(x,t) \\
&= \frac{1}{1+f'(y,t)dt}\left\langle \rho(y-f(y,t)dt-\xi,t)\right\rangle_\xi \\
&= \frac{1}{1+f'(y,t)dt}\left\langle \rho(y,t) - (f(y,t)dt+\xi)\nabla\rho(y,t) + \frac{1}{2}(f(y,t)+\xi)^2\nabla^2\rho(y,t)\right\rangle_\xi \\
&= \rho(y,t) - \nabla(f(y,t)\rho(y,t))dt + \frac{1}{2}\nu\nabla^2\rho(y,t)dt
\end{aligned}$$

Thus,

$$\partial_t\rho(x,t) = -\nabla(f(x,t)\rho(x,t)) + \frac{1}{2}\nu\nabla^2\rho(x,t), \qquad \rho(x,0) = \delta(x-x_0)$$

# Example: Brownian motion

$$
\begin{aligned}
dx &= d\xi \qquad \left\langle d\xi^2 \right\rangle = \nu dt \\
\rho(x,t) &= p(x,t|x_0,0) = \frac{1}{\sqrt{2\pi\nu t}} \exp\left(-\frac{(x-x_0)^2}{2\nu t}\right) \\
\psi(x,t) &= p(z,T|x,t) = \frac{1}{\sqrt{2\pi\nu(T-t)}} \exp\left(-\frac{(x-z)^2}{2\nu(T-t)}\right)
\end{aligned}
$$

# Forward and backward drift

For

$$dx = f(x,t)dt + \xi$$

The *expected forward drift* is

$$\langle dx \rangle = f(x,t)dt$$

The *expected backward drift* given $x(t+dt) = y$ can be computed using Bayes' rule:

$$\begin{aligned} p(y, t-dt|x,t) &= \frac{p(x,t|y,t-dt)\rho(y,t-dt)}{\rho(x,t)} \\ p(x,t|y,t-dt) &= \langle \delta(x-y-f(y,t-dt)dt-\xi) \rangle_\xi \end{aligned}$$

$$\begin{aligned}
&\langle x(t) - y(t-dt)\rangle_{x(t)=x} = \int dy(x-y)p(y,t-dt|x,t) \\
= &\int dy(x-y)\left\langle \delta(x-y-f(y,t-dt)dt-\xi)\right\rangle \frac{\rho(y,t-dt)}{\rho(x,t)} \\
= &\frac{1}{\rho(x,t)}\left\langle \frac{1}{1+f'(x,t)dt}(f(x,t)dt+\xi)\rho(x-f(x,t)dt-\xi,t-dt)\right\rangle + \mathcal{O}(dt^2) \\
= &\frac{1}{\rho(x,t)}\frac{1}{1+f'(x,t)dt}\left\langle (f(x,t)dt+\xi)(\rho(x,t)-\xi\rho'(x,t)\right\rangle + \mathcal{O}(dt^2) \\
= &f(x,t)dt - \nu\nabla\log\rho(x,t)dt + \mathcal{O}(dt^2) \equiv \tilde{f}(x,t)dt
\end{aligned}$$

We see that the forward and backward drifts are different: given that we are at time $t$ at location $x$ the expected future drift is given by $f(x,t)$. The expected past drift into $x$ is given by $\tilde{f}(x,t) = f(x,t) - \nu\nabla\log\rho(x,t)$.

# Example: Brownian motion

$$
\begin{aligned}
dx &= d\xi \qquad x(0) = 0 \quad \left\langle d\xi^2 \right\rangle = \nu dt \\
\rho(x,t) &= \frac{1}{\sqrt{2\pi\nu t}} \exp\left(-\frac{x^2}{2\nu t}\right) \\
f(x,t) &= 0 \\
\tilde{f}(x,t) &= -\frac{x}{t}
\end{aligned}
$$

# Lecture 3:

# Stochastic optimal control

Consider a stochastic dynamical system

$$dx = f(t, x, u)dt + d\xi$$

$d\xi$ Gaussian noise $\langle d\xi_i d\xi_j \rangle = \nu_{ij}(t, x, u)dt$.

The cost becomes an expectation:

$$C(t, x, u(t \to T)) \quad = \quad \left\langle \phi(x(T)) + \int_t^T d\tau R(t, x(t), u(t)) \right\rangle$$

over all stochastic trajectories starting at $x$ with control path $u(t \to T)$.

Note, that $u(t)$ as part of $u(t \to T)$ is used at time $t$. Next move to $x + dx$ and repeat the optimization.

# Stochastic optimal control

We obtain the Bellman recursion

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

Thus,

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

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

# Linear Quadratic control

The dynamics is linear

$$dx = [A(t)x + B(t)u + b(t)]dt + \sum_{j=1}^{m}(C_j(t)x + D_j(t)u + \sigma_j(t))d\xi_j, \quad \left\langle d\xi_j d\xi_{j'} \right\rangle = \delta$$

The cost function is quadratic

$$\begin{aligned} \phi(x) &= \frac{1}{2}x^T G x \\ R(x,u,t) &= \frac{1}{2}x^T Q(t)x + u^T S(t)x + \frac{1}{2}u^T R(t)u \end{aligned}$$

In this case the optimal cost-to-go is quadratic in $x$:

$$\begin{aligned} J(t,x) &= \frac{1}{2}x^T P(t)x + \alpha^T(t)x + \beta(t) \\ u(t) &= -\Psi(t)x(t) - \psi(t) \end{aligned}$$

Substitution in the HJB equation yields ODEs for $P, \alpha, \beta$:

$$
\begin{aligned}
-\dot{P} &= PA + A^T P + \sum_{j=1}^{m} C_j^T P C_j + Q - \hat{S}^T \hat{R}^{-1} \hat{S} \\
-\dot{\alpha} &= [A - B\hat{R}^{-1}\hat{S}]^T \alpha + \sum_{j=1}^{m} [C_j - D_j \hat{R}^{-1}\hat{S}]^T P \sigma_j + Pb \\
\dot{\beta} &= \frac{1}{2} \left| \sqrt{\hat{R}} \psi \right|^2 - \alpha^T b - \frac{1}{2} \sum_{j=1}^{m} \sigma_j^T P \sigma_j \\
\hat{R} &= R + \sum_{j=1}^{m} D_j^T P D_j \\
\hat{S} &= B^T P + S + \sum_{j=1}^{m} D_j^T P C_j \\
\Psi &= \hat{R}^{-1} \hat{S} \\
\psi &= \hat{R}^{-1} (B^T \alpha + \sum_{j=1}^{m} D_j^T P \sigma_j)
\end{aligned}
$$

with $P(t_f) = G$ and $\alpha(t_f) = \beta(t_f) = 0$.

# Example

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 $R(x,u) = \frac{1}{2}(Qx^2 + Ru^2)$.

The Ricatti equations reduce to

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

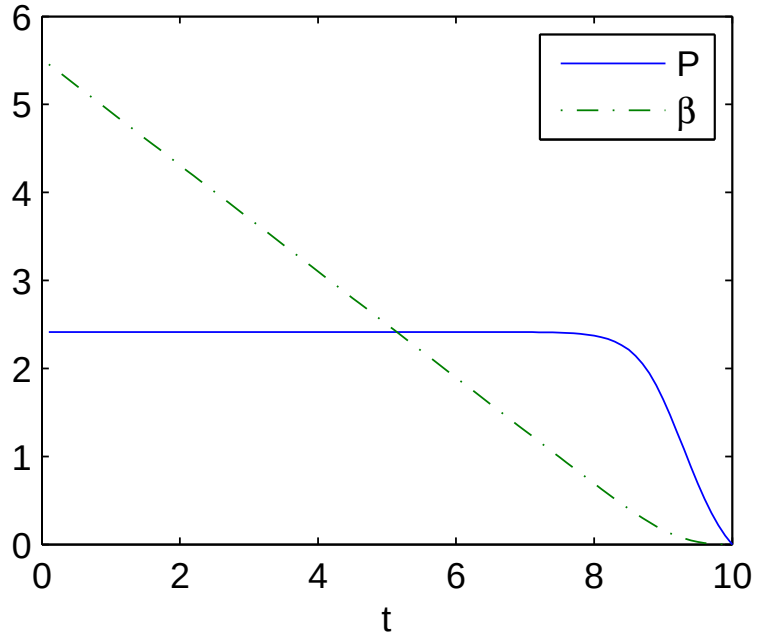
with $P(T) = \alpha(T) = \beta(T) = 0$ and

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

The solution is

$$P(t) = R\frac{\exp(2\sqrt{1+R^{-1}Q}(T-t))-1}{\frac{1}{1+\sqrt{1+R^{-1}Q}}\exp(2\sqrt{1+R^{-1}Q}(T-t)) - \frac{1}{1-\sqrt{1+R^{-1}Q}}}$$

The optimal control is $u(x,t) = -R^{-1}P(t)x$.



# Comments

Note, that in the last example the optimal control is independent of $\nu$, i.e. optimal stochastic control equals optimal deterministic control.

In general:

- If $C_j = D_j = 0$ (only 'additive noise') $\dot{P}, \dot{\alpha}$ independent of noise $\sigma$, $\dot{\beta}$ depends on $\sigma$, but control independent of $\beta$. Thus control independent of $\sigma$ (certainty equivalence)
- If $C_j \neq 0$ or $D_j \neq 0$, control depends on $C_j, D_j, \sigma_j$ (no certainty equivalence)

# Example: Portfolio selection

[2] Consider a market with $p$ stocks and one bond. The bond price process is subject ot the following deterministic ordinary differential equation:

$$dP_0(t) = r(t)P_0(t)dt, \quad P_0(0) = p_0 > 0 \tag{1}$$

The other assets have price processes $P_i(t), i = 1, \ldots, p$ satisfying stochastic differential equations

$$dP_i(t) = P_i(t)\left(b_i(t)dt + \sum_{j=1}^{m} \sigma_{ij}(t)d\xi_j(t)\right), \quad P_i(0) = p_i > 0 \tag{2}$$

Consider an investor whose total wealth at time $t$ is denoted by $x(t)$

$$x(t) = \sum_{i=0}^{p} N_i(t)P_i(t) \tag{3}$$

[2] This section is from [1] section 6.8 (pg. 335).

with $N_i$ the number of stocks/bond of type $i$. Then

$$dx(t) = \left( r(t)x(t) + \sum_{i=1}^{p} (b_i(t) - r(t))u_i(t) \right) dt + \sum_{i=1}^{p} \sum_{j=1}^{m} \sigma_{ij}(t)u_i(t)d\xi_j(t) \quad (4)$$

with $u_i(t) = N_i(t)P_i(t), i = 1, \ldots, p$ the *portfolio* of the investor.

The objective of the investor is to maximize the mean terminal wealth $\langle x(t_f) \rangle$ and minimize at the same time the variance

$$\Sigma^2 = \left\langle x(t_f)^2 \right\rangle - \langle x(t_f) \rangle^2$$

This is a multi-objective optimization problem with an efficient frontier of optimal solutions: for each given mean there is a minimial variance.

These pairs can be found by minimizing the single objective criterion

$$\mu\Sigma^2 - \langle x(t_f) \rangle \quad (5)$$

for different values of the weighting factor $\mu$.

This objective, however, is not an expectation value of some stochastic quantity due to the $\langle\cdot\rangle^2$ term. Consider a slightly different problem, minimizing the objective

$$\left\langle \mu x(t_f)^2 - \lambda x(t_f) \right\rangle \tag{6}$$

which is of the standard stochastic optimization form. One can show that one can construct a solution of Problem 5 by solving problem 6 for suitable $\lambda(\mu)$. [3]

Our goal is thus to minimize eq. 6 subject to the stochastic dynamics eq. 4.

This is an LQ problem. The solution is computed from the Ricatti equations

$$u_i(x,t) = \psi_i(t)x + \phi_i(t)$$

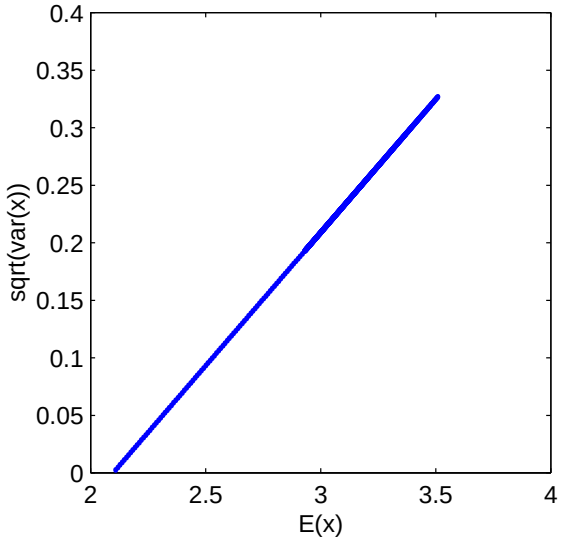
As an example we consider the simplest possible case: $p = m = 1$ and $r, b, \sigma$ independent of time.

[3] and finding $\lambda$ from

$$\lambda = 1 + 2\mu \left\langle x(t_f) \right\rangle (\lambda, \mu)$$

([1] Theorem 8.2 pg. 338)

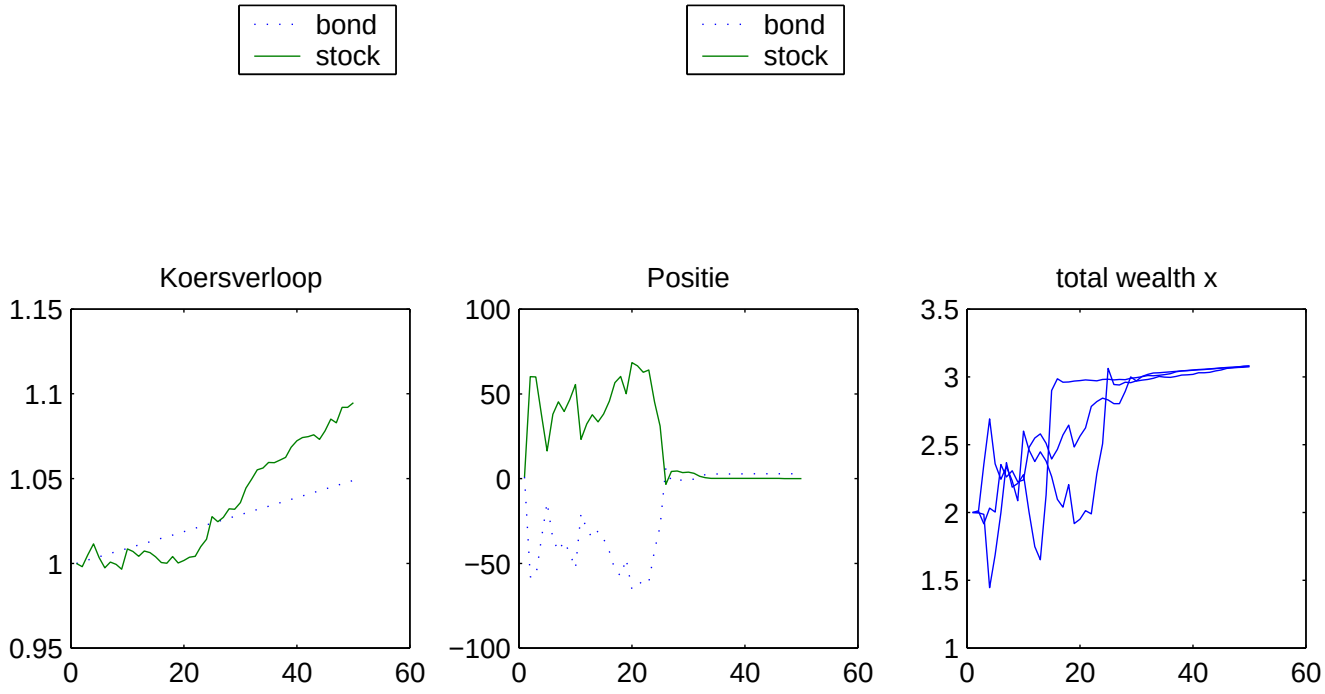
# Efficient boundary




Parameter values are: $p = m = 1$. Trading period is one year weekly. annual bond rate 5 % ($r = 0.0009758$), annual expected stock rate is 10 % ($b = 0.0019$), volatility $\sigma = 2b$. $x_0 = 2$. Shows var $x$ versus $\langle x \rangle$ scatter plot for various values of $\mu$. Small $\mu$ corresponds to risky investments with high expected return and large fluctuation. $\mu \longrightarrow \infty$ corresponds to riskless investment in bond only and a return of 5 %.

$\mu = 10$ corresponds to $\langle x \rangle = 3$ and $\sqrt{var} = 0.2$.

# Making money




Simulation of optimal control with $\mu = 10$, The optimal strategy is to borrow many stocks and sell them as soon as the objective is achieved.

Indeed, $\langle x \rangle = 3$ as expected. The strategy to get at this 50 % increase in wealth is to buy many stocks and hope they will give the expected wealth increase. As soon as this occurs, all stocks are sold and the money is put in the bank.

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

For instance:

- For LQ control problems the optimal control computation is equivalent to 'Kalman smoothing'.

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

Variational inference:

$p(x_{1:n}) = \pi(x_{1:n})/Z$ is a probability distribution, compute

$$p(x_1) \quad = \quad \sum_{x_{2:n}} p(x_{1:n})$$

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

Variational inference:

$p(x_{1:n}) = \pi(x_{1:n})/Z$ is a probability distribution, compute

$$p(x_1) \quad = \quad \sum_{x_{2:n}} p(x_{1:n})$$

Define free energy

$$F(q) \quad = \quad \sum_{x_{1:n}} q(x_{1:n}) \log \frac{q(x_{1:n})}{\pi(x_{1:n})}$$

$F$ is minimized by $q = p$.

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

Variational inference:

$p(x_{1:n}) = \pi(x_{1:n})/Z$ is a probability distribution, compute

$$p(x_1) = \sum_{x_{2:n}} p(x_1, x_{2:n})$$

Define free energy

$$F(q) = \sum_{x_{1:n}} q(x_{1:n}) \log \frac{q(x_{1:n})}{\pi(x_{1:n})}$$

$F$ is minimized by $q = p$.

Restrict minimization to simple distributions $q(x_{1:n}) = q_1(x_1) \dots q_n(x_n)$ and minimize

$$p(x_1) \approx q_1(x_1)$$

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

Efficient inference:

- Variational inference, TAP
- Belief propagation, EP, Cluster Variation Method, Survey propagation
- convex relaxations
- Monte Carlo Sampling

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

In particular:

- Consider a class of control problems for which the Bellman equation can be transformed in a linear pde (using a log transform)
- 'Solve' as a Feynman-Kac path integral

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient state-of-the-art inference methods

The log transform first used in QM:

$$\hbar i \partial_t \Psi = H \Psi \qquad H(x,t) = V(x,t) - \frac{\hbar^2}{2} \partial_x^2$$

Write

$$\Psi = \sqrt{\rho} \exp\left(i \frac{S}{\hbar}\right)$$

then

$$\begin{aligned} -\partial_t S &= \frac{1}{2}(\nabla_x S)^2 - \frac{1}{2}\hbar^2 \frac{\partial_x^2 \sqrt{\rho}}{\sqrt{\rho}} + V \\ -\partial_t \rho &= \nabla_x(\rho \nabla_x S) \end{aligned}$$

Later used in Burgers Equation, and by Fleming and Mitter for control.

# Link between control and inference

General idea:

- Express the control problem as an inference problem
- Use efficient approximate inference methods

In particular:

- Consider a class of control problems for which the Bellman equation looks like the Mandelung equation
- Use the log transform to convert it into a Schrödinger-like backward equation
- Identify this equation as a Kolmogorov backward equation.
- Identify the corresponding forward diffusion process

# Path integral control

$$
\begin{aligned}
dx_i &= f_i(x,t)dt + \sum_j g_{ia}(x,t)(u_a dt + d\xi_a) \\
C(t,x,u(t\to T)) &= \left\langle \phi(x(T)) + \int_t^T ds V(x,t) + \frac{1}{2}\sum_{ab} R_{ab}u_a u_b \right\rangle
\end{aligned}
$$

with $\langle d\xi_a d\xi_b \rangle = \nu_{ab} dt$.

The cost is an expectation over all stochastic trajectories starting at $x$ with control path $u(t \to T)$.

The stochastic HJB equation becomes

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

which we need to solve with end boundary condition $J(x, t_f) = \phi(x)$.

# Path integral control

Minimization wrt $u$ yields: [4]

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

(our 'Mandelung equation')

Define $\psi(x,t)$ through $J(x,t) = -\lambda \log \psi(x,t)$ and impose a relation between $R$ and $\nu$:

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

with $\lambda$ a positive number.

[4] $u_a = -\sum_b \left(R^{-1}\right)_{ab} g_{ib}(x,t) \frac{\partial J(x,t)}{\partial x_i}$

# The relation $R = \lambda \nu^{-1}$

$$
\begin{aligned}
dx_i &= f_i(x)dt + \sum_a g_{ia}(x)(u_a dt + d\xi_a) \\
C &= \left\langle \phi(x(T)) + \int_t^T ds V(x) + \frac{1}{2}\sum_{ab} R_{ab} u_a u_b \right\rangle
\end{aligned}
$$

Noise and control act in the same sub-space. Directions where noise is large, control is cheap and visa versa.

# The relation $R = \lambda\nu^{-1}$

$$
\begin{aligned}
dx_i &= f_i(x)dt + \sum_a g_{ia}(x)(u_a dt + d\xi_a) \\
C &= \left\langle \phi(x(T)) + \int_t^T dsV(x) + \frac{1}{2}\sum_{ab} R_{ab}u_a u_b \right\rangle
\end{aligned}
$$

Noise and control act in the same sub-space. Directions where noise is large, control is cheap and visa versa.

Can be alternatively understood as a KL divergence between controlled and uncontrolled trajectories:

$$
\sum_\tau p(\tau|u)\log\frac{p(\tau|u)}{p(\tau|0)} = \int_0^T dt\frac{1}{2}u^T\nu^{-1}u
$$

$\lambda$ plays the role of temperature.

# Path integral control

Then the HJB becomes *linear* in $\psi$

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

with end condition $\psi(x, T) = \exp(-\phi(x)/\lambda)$ (our Kolmogorov backward equation) [5]

[5] We sketch the derivation for $g = 1$.

$$-\frac{1}{2} (\nabla J)^T R^{-1} (\nabla J) + \frac{1}{2} \mathrm{Tr} \left( \nu \nabla^2 J \right) = -\frac{1}{2} \sum_{ij} \nabla_i J R^{-1}_{ij} \nabla_j J + \frac{1}{2} \lambda \sum_{ij} R^{-1}_{ij} \nabla_{ij} J$$

$$= \frac{1}{2} \sum_{ij} R^{-1}_{ij} \left( -\nabla_i J \nabla_j J + \lambda \nabla_{ij} J \right)$$

$$= \frac{1}{2} \sum_{ij} R^{-1}_{ij} \left( -\lambda^2 \frac{1}{\psi} \nabla_{ij} \psi \right)$$

since

$$-\nabla_i J \nabla_j J \quad = \quad -\lambda^2 \frac{1}{\psi^2} \nabla_i \psi \nabla_j \psi$$

# Path integral control

The linearity allows us to revers the direction of time.

We identify $\psi(x,t) \propto p(z,T|x,t)$, then the Bellman equation

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

can be interpreted as a Kolmogorov backward equation for the process

$$\begin{aligned} dx_i &= f_i(x,t)dt + \sum_a g_{ia}(x,t) d\xi_a \\ x(t) &= \dagger \quad \text{with probability} \quad V(x,t)dt/\lambda \\ x(T) &= \dagger \quad \text{with probability} \quad \phi(x)/\lambda \end{aligned}$$

---

$$\nabla_{ij} J = -\lambda \nabla_i \nabla_j \log \psi = -\lambda \nabla_i \left( \frac{1}{\psi} \nabla_j \psi \right) = \lambda \frac{1}{\psi^2} \nabla_i \psi \nabla_j \psi - \lambda \frac{1}{\psi} \nabla_{ij} \psi$$

# Path integral control

The correspondong forward equation is

$$\partial_t \rho = -\frac{V}{\lambda}\rho - \nabla(f\rho) + \frac{1}{2}\mathrm{Tr}\nabla^2 g\nu g^T \rho$$

with $\rho(x,t) = p(x,t|z,0)$ and $\rho(x,0) = \delta(x-z)$.

# Feynman-Kac formula

Denote $Q(\tau|x,s)$ the distribution over uncontrolled trajectories that start at $x,t$:

$$dx = f(x,t)dt + g(x,t)d\xi$$

with $\tau$ a trajectory $x(t \to T)$. Then

$$\psi(x,t) = \int dQ(\tau|x,t) \exp\left(-\frac{S(\tau)}{\lambda}\right)$$

$$S(\tau) = \phi(x(T)) + \int_t^T dsV(x(s),s)$$

$\psi$ can be computed by forward sampling the uncontrolled process.

# Posterior distribution over optimal trajectories

$\psi(x,t)$ can be interpreted as a partition sum for the distribution over paths under optimal control:

$$P(\tau|x,t) = \frac{1}{\psi(x,t)} Q(\tau|x,t) \exp\left(-\frac{S(\tau)}{\lambda}\right)$$

The optimal cost-to-go is a free energy:

$$J(x,t) = -\lambda \log \int dQ(\tau|x,t) \exp\left(-\frac{1}{\lambda} S(\tau)\right)$$

The optimal control is an expectation wrt $P$:

$$u(x,t)dt = -R^{-1} g^T(x,t) \nabla_x J(x,t) dt = \int dP(\tau) d\xi(\tau) = \langle d\xi \rangle_P$$

# Recap

Control problem:

$$dx = f dt + g(u dt + d\xi) \qquad C = \left\langle \phi + \int_t^T V + \frac{1}{2} u^T R u \right\rangle \qquad R = \lambda \nu^{-1}$$

HJB is linear:

$$\partial_t \psi = H \psi \qquad J = -\lambda \log \psi$$

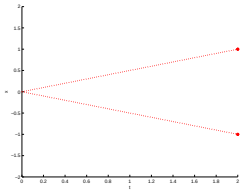
Solution is given by Feynman-Kac formula: $\psi = \int dQ(\tau) \exp\left(-\frac{S(\tau)}{\lambda}\right)$.
$Q$ distribution over uncontrolled dynamics ($u = 0$).

Posterior distribution over optimal controlled trajectories: $P(\tau) = \frac{1}{\psi} Q(\tau) \exp\left(-\frac{S(\tau)}{\lambda}\right)$.

Optimal control is expectation value: $u dt = \langle d\xi \rangle_P$.

# Delayed choice

$$dx = u_t dt + d\xi_t \qquad \left\langle \xi_t^2 \right\rangle = \nu dt$$



$V = 0$, path cost is $\frac{1}{2}u^2$ and end cost $\phi(z = \pm 1) = 0, \phi(z) = \infty$ else encodes two targets at $z = \pm 1$ at $t = T$.

PI recipe:

1.

$$\begin{aligned}
\psi(x,t) &= \int dQ(\tau|x,t) \exp(-S(\tau)/\lambda) \\
S(\tau) &= \phi(x(T)) \\
\psi(x,t) &= \int dz q(z,T|x,t) \exp(-\phi(z)/\lambda) = q(1,T|x,t) + q(-1,T|x,t) \\
q(z,T|x,t) &= \mathcal{N}(z|x,\nu(T-t))
\end{aligned}$$

2. Compute

$$J(x,t) \quad = \quad -\lambda \log \psi(x,t) = \frac{1}{T-t}\left(\frac{1}{2}x^2 - \nu(T-t)\log 2\cosh\frac{x}{\nu(T-t)}\right)$$

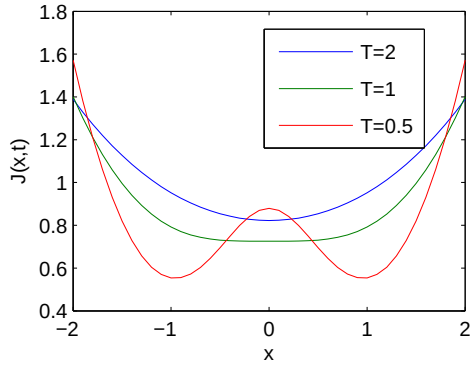
3.

$$u(x,t) \quad = \quad -\nabla J(x,t) = \frac{1}{T-t}\left(\tanh\frac{x}{\nu(T-t)} - x\right)$$

# Delayed choice

$$dx = udt + d\xi \qquad \left\langle \xi^2 \right\rangle = \nu dt$$

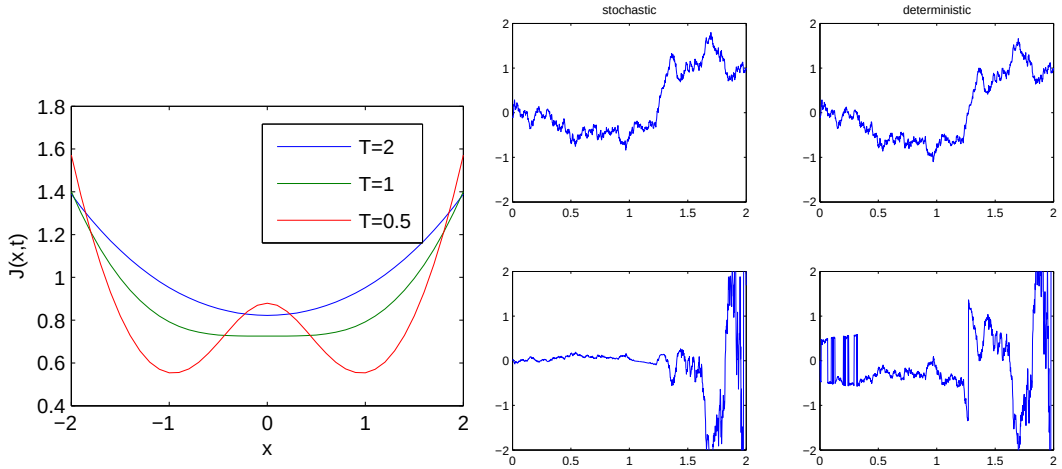
$V = 0$, path cost is $\frac{1}{2}u^2$, $\phi(x = \pm 1) = 0$ and $\phi(x) = \infty$, else.



# Delayed choice

$$dx = udt + d\xi \qquad \left\langle \xi^2 \right\rangle = \nu dt$$

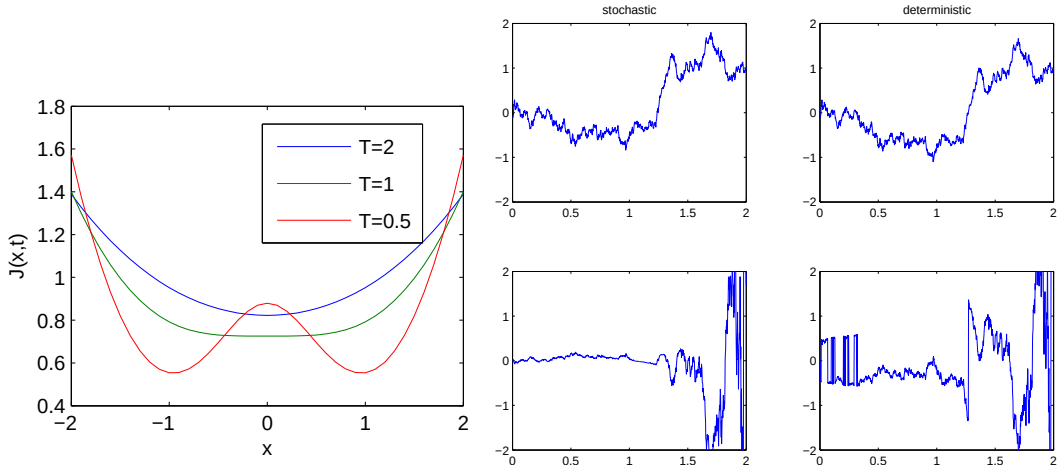
$V = 0$, path cost is $\frac{1}{2}u^2$, $\phi(x = \pm 1) = 0$ and $\phi(x) = \infty$, else.



# Delayed choice

$$dx = udt + d\xi \qquad \left\langle \xi^2 \right\rangle = \nu dt$$

$V = 0$, path cost is $\frac{1}{2}u^2$, $\phi(x = \pm 1) = 0$ and $\phi(x) = \infty$, else.




"When the future is uncertain, delay your decisions."

# Estimating optimal control by sampling

For given $x, t$, the optimal control is given by

$$udt = \int dP(\tau) d\xi(\tau) = \frac{\int dQ(\tau) \exp(-S(\tau)/\lambda) d\xi(\tau)}{\int dQ(\tau) \exp(-S(\tau)/\lambda)}$$

We generate $N$ trajectories $x^\mu_{t:T}$ starting at $x, t$ with initial noise $d\xi^\mu$. Define

$$\begin{aligned} S^\mu &= \sum_{s=t}^{T} V(x^\mu_s, s)dt + \phi(x^\mu_T) \\ udt &= \frac{\sum_\mu d\xi^\mu \exp(-S^\mu/\lambda)}{\sum_\mu \exp(-S^\mu/\lambda)} \end{aligned}$$

Unbiased, but inefficient.

# Importance sampling

Efficiency may be improved by sampling with $u \neq 0$.

$$\psi(x,t) = \int dQ(\tau) \exp(-S(\tau)/\lambda) = \int dQ'(\tau) \frac{dQ(\tau)}{dQ'(\tau)} \exp(-S(\tau)/\lambda)$$

with $Q'(\tau|x,t)$ from the stochastic process

$$dx = f(x,t)dt + g(x,t)(\hat{u}(x,t)dt + d\xi)$$

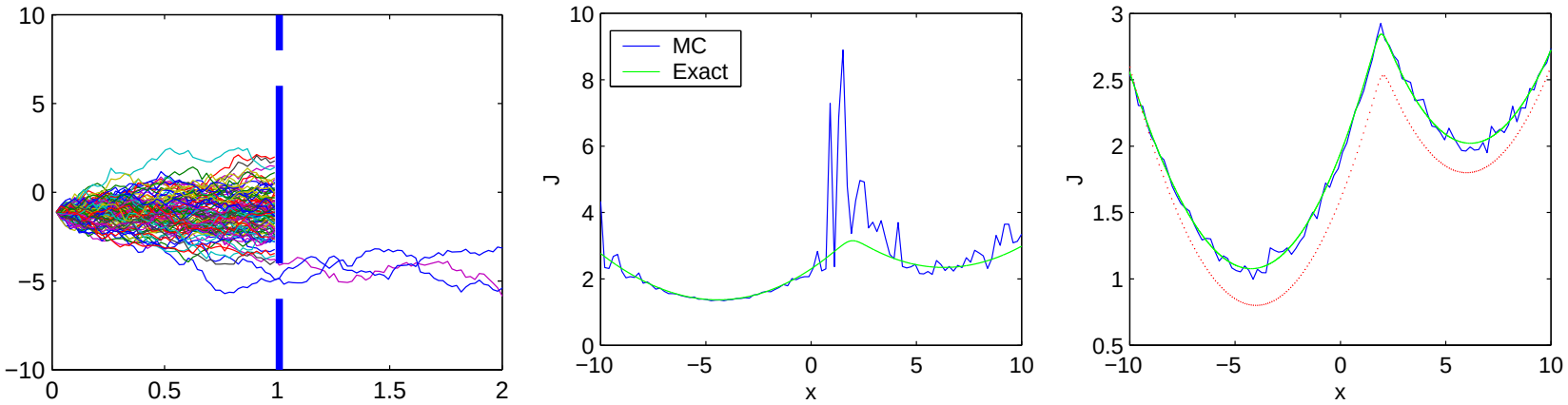
# Importance sampling

Efficiency may be improved by sampling with $u \neq 0$.

$$\psi(x,t) = \int dQ(\tau) \exp(-S(\tau)/\lambda) = \int dQ'(\tau) \frac{dQ(\tau)}{dQ'(\tau)} \exp(-S(\tau)/\lambda)$$

with $Q'(\tau|x,t)$ from the stochastic process

$$dx = f(x,t)dt + g(x,t)(\hat{u}(x,t)dt + d\xi)$$

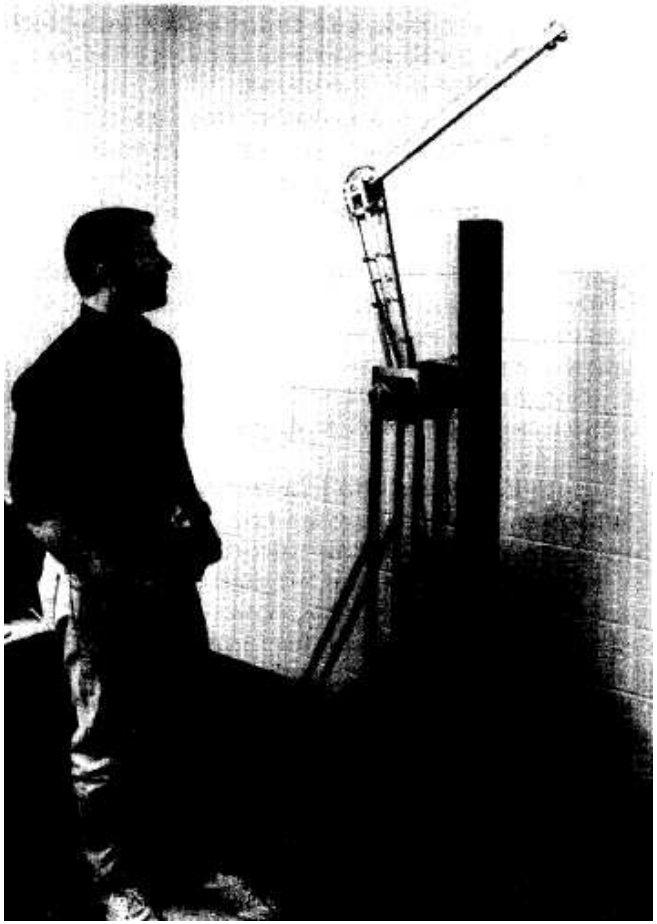


# How to control a device?

Plant is unknown

Exploration of state space

Motor babbling in infants



# How to control a device?

Plant is unknown

Exploration of state space

Motor babbling in infants

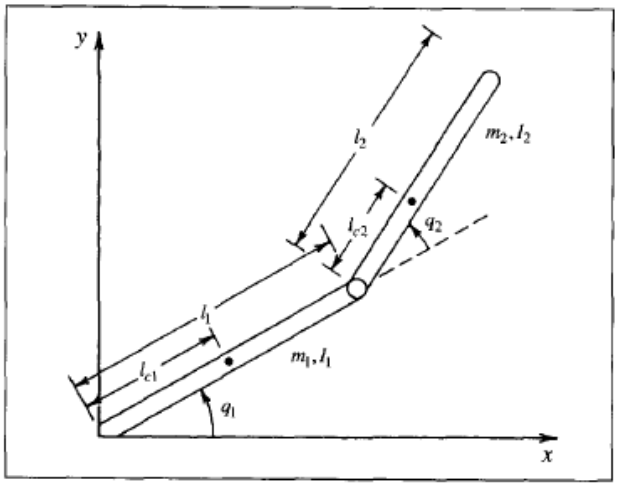



*Fig. 1. The Acrobot.*

$$d_{11}\ddot{q}_1 + d_{12}\ddot{q}_2 + h_1 + \phi_1 = 0 \quad (1)$$

$$d_{21}\ddot{q}_1 + d_{22}\ddot{q}_2 + h_2 + \phi_2 = \tau, \quad (2)$$

where

$$d_{11} = m_1 l_{c1}^2 + m_2(l_1^2 + l_{c2}^2 + 2 l_1 l_{c2}\cos(q_2)) + I_1 + I_2$$
$$d_{22} = m_2 l_{c2}^2 + I_2$$
$$d_{12} = m_2(l_{c2}^2 + l_1 l_{c2}\cos(q_2)) + I_2$$
$$d_{21} = m_2(l_{c2}^2 + l_1 l_{c2}\cos(q_2)) + I_2$$
$$h_1 = -m_2 l_1 l_{c2}\sin(q_2)\dot{q}_2^2 - 2 m_2 l_1 l_{c2}\sin(q_2)\dot{q}_2\dot{q}_1$$
$$h_2 = m_2 l_1 l_{c2}\sin(q_2)\dot{q}_1^2$$
$$\phi_1 = (m_1 l_{c1} + m_2 l_1)g\cos(q_1) + m_2 l_{c2} g\cos(q_1 + q_2)$$
$$\phi_2 = m_2 l_{c2} g\cos(q_1 + q_2).$$

Problem for brains and for robots

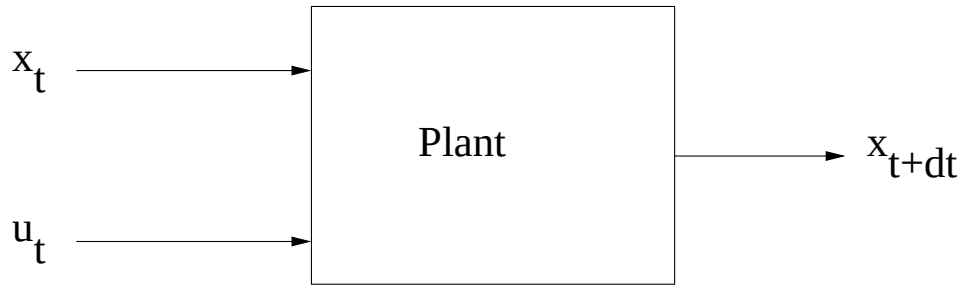
# Control of a deterministic unknown plant

We consider a deterministic control problem of the form

$$\begin{aligned} dx_i &= f_i(x,t)dt + \sum_a g_{ia}(x,t)u_a dt \\ C &= \int_0^T \frac{1}{2}u^T R u + V(x,t) \end{aligned}$$

the problem is to compute the optimal control law $u_a(x,t)$ from a sequence of states that we generate with some chosen control

$$u^\mu_{0:T}, x^\mu_{0:T}, \qquad \mu = 1, \ldots, N$$



Suppose that we choose random controls from a Gaussian distribution: $u_a dt = d\xi_a, \nu = \lambda R^{-1}$. The dynamics becomes

$$dx_i = f_i(x,t)dt + \sum_a g_{ia}(x,t)d\xi_a$$

the uncontrolled dynamics of the stochastic control

$$dx_i = f_i(x,t)dt + \sum_a g_{ia}(x,t)(u_a dt + d\xi_a)$$

$$C = \left\langle \int_0^T \frac{1}{2} u^T R u + V(x,t) \right\rangle$$

which is equivalent to the original control problem when $\lambda \to 0$.

# Acrobot

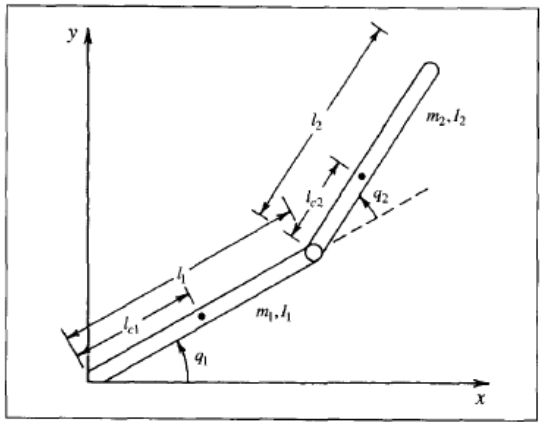



*Fig. 1. The Acrobot.*

$$d_{11}\ddot{q}_1 + d_{12}\ddot{q}_2 + h_1 + \phi_1 = 0 \quad (1)$$

$$d_{21}\ddot{q}_1 + d_{22}\ddot{q}_2 + h_2 + \phi_2 = \tau, \quad (2)$$

where

$$d_{11} = m_1 l_{c1}^2 + m_2(l_1^2 + l_{c2}^2 + 2 l_1 l_{c2}\cos(q_2)) + I_1 + I_2$$
$$d_{22} = m_2 l_{c2}^2 + I_2$$
$$d_{12} = m_2(l_{c2}^2 + l_1 l_{c2}\cos(q_2)) + I_2$$
$$d_{21} = m_2(l_{c2}^2 + l_1 l_{c2}\cos(q_2)) + I_2$$
$$h_1 = -m_2 l_1 l_{c2}\sin(q_2)\dot{q}_2^2 - 2 m_2 l_1 l_{c2}\sin(q_2)\dot{q}_2\dot{q}_1$$
$$h_2 = m_2 l_1 l_{c2}\sin(q_2)\dot{q}_1^2$$
$$\phi_1 = (m_1 l_{c1} + m_2 l_1)g\cos(q_1) + m_2 l_{c2} g\cos(q_1 + q_2)$$
$$\phi_2 = m_2 l_{c2} g\cos(q_1 + q_2).$$

$q_1(0) = q_2(0) = -\pi/2$, $\dot{q}_1(0) = \dot{q}_2(0) = 0$, maximize final height

$$H = l_1 \sin q_1(T) + l_2 \sin q_2(T)$$

# Acrobot

$$d_{11}(q)\ddot{q}_1 + d_{12}(q)\ddot{q}_2 + h_1(q,\dot{q}) + \phi_1(q) = 0$$
$$d_{21}(q)\ddot{q}_1 + d_{22}\ddot{q}_2 + h_2(q,\dot{q}) + \phi_2(q) = u$$

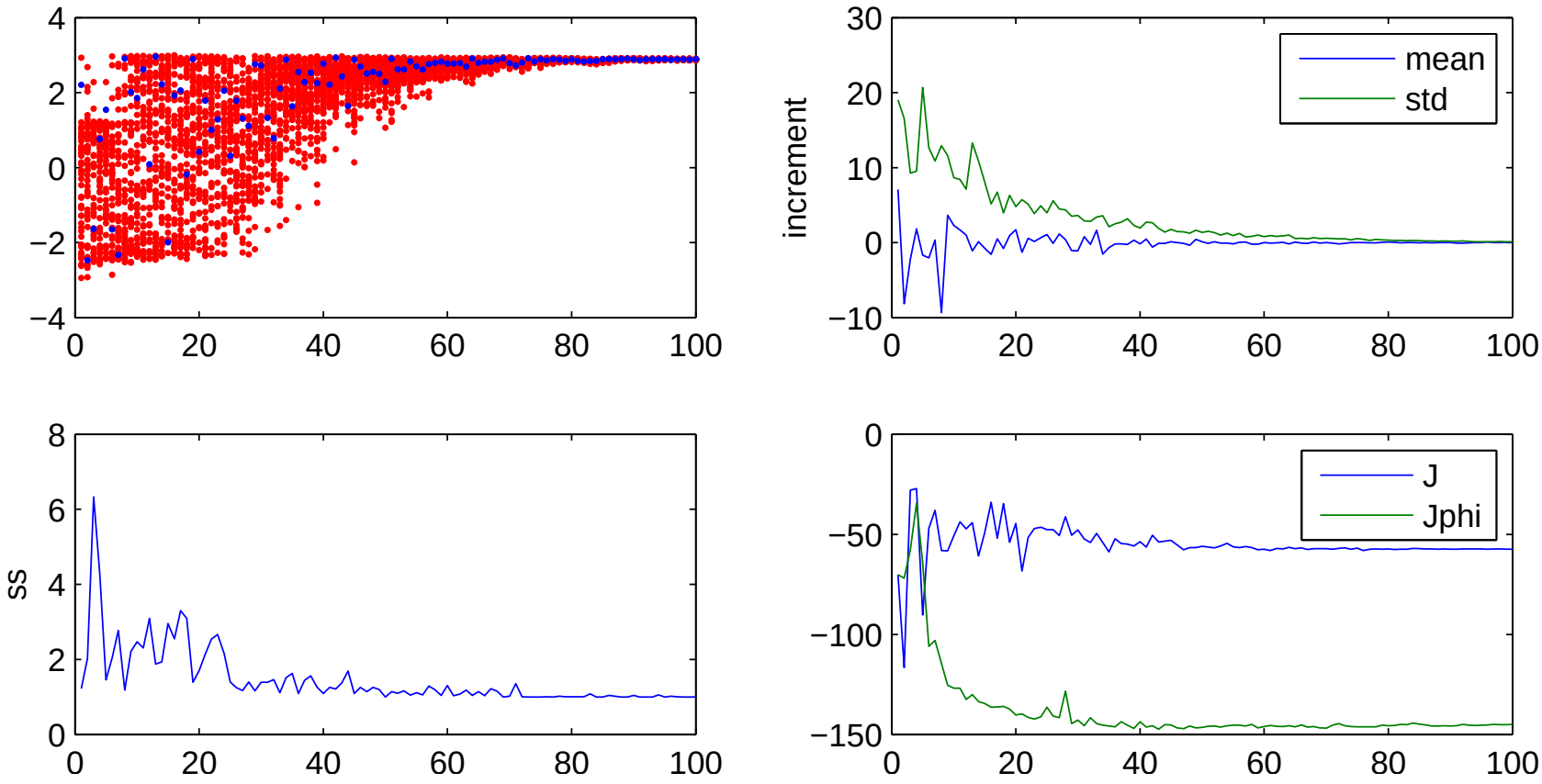
We can write these equations in standard form

$$dx_i = f_i(x)dt + g_i(x)udt$$

with $x_1 = q_1, x_2 = q_2, x_3 = \dot{q}_1, x_4 = \dot{q}_2$ and

$$\begin{aligned} f_1(x) &= x_3 & g_1(x) &= 0 \\ f_2(x) &= x_4 & g_2(x) &= 0 \\ f_3(x) &= \frac{-d_{22}(h_1+\phi_1)+d_{12}(h_2+\phi_2)}{D} & g_3(x) &= -\frac{d_{12}}{D} \\ f_4(x) &= \frac{d_{12}(h_1+\phi_1)-d_{11}(h_2+\phi_2)}{D} & g_4(x) &= \frac{d_{11}}{D} \end{aligned}$$

# Acrobot

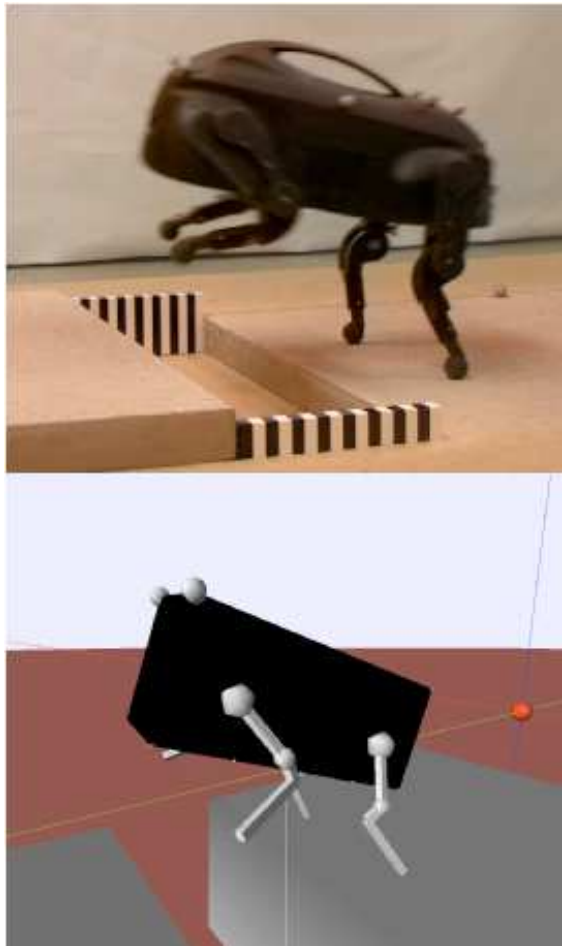



100 iterations. At each iteration $50$ stochastic trajectories were generated. The new control was computed from a deterministic trajectory. Noise was lowered at each iteration. Top left: final height for each stochastic trajectory for each iteration (red) and for each deterministic solution (blue).

# Acrobot

Result after 100 trials

# Application in robotics



Compares favorably to state-of-the-art RL methods (Theodorou et al. 2010-2012)

# KL control theory

$x$ denotes state of the agent and $x_{1:T}$ is a path through state space from time $t = 1$ to $T$.

$q(x_{1:T}|x_0)$ denotes a probability distribution over possible future trajectories given that the agent at time $t = 0$ is is state $x_0$, with

$$q(x_{1:T}|x_0) = \prod_{t=0}^{T} q(x_{t+1}|x_t)$$

$q(x_{t+1}|x_t)$ implements the allowed moves.

$R(x_{1:T}) = \sum_{t=1}^{T} R(x_t)$ is the total cost when following path $x_{1:T}$.

The KL control problem is to find the probability distribution $p(x_{1:T}|x_0)$ that minimizes

$$C(p|x_0) \quad = \quad \sum_{x_{1:T}} p(x_{1:T}|x_0) \left( \log \frac{p(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} + R(x_{1:T}) \right) = KL(p||q) + \langle R \rangle_p$$

# KL control theory

$p(x_{1:T}|x_0)$ and $q(x_{1:T}|x_0)$ distributions over trajectories.

Given $q$, find $p$ that minimizes

$$C(p|x_0) = KL(p||q) - \langle R \rangle_p$$

The solution and the optimal control cost are

$$\begin{aligned} p(x_{1:T}|x_0) &= \frac{1}{Z(x_0)} q(x_{1:T}|x_0) \exp\left(R(x_{1:T})\right) \\ C &= -\log Z(x_0) \\ Z(x_0) &= \sum_{x_{1:T}} q(x_{1:T}|x_0) \exp\left(R(x_{1:T})\right) \end{aligned}$$

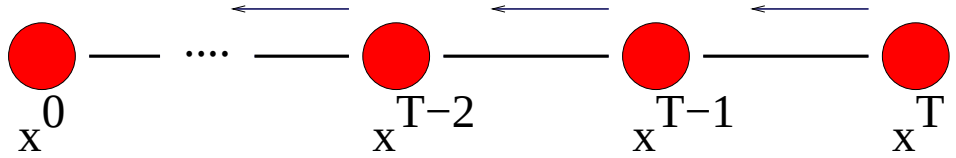
NB: $Z(x_0)$ is an integral over paths.

# KL control theory

The optimal control at time $t = 0$ is given by

$$p(x_1|x_0) \quad = \quad \sum_{x_{2:T}} p(x_{2:T}|x_0) \propto q(x_1|x_0) \exp(R(x_1)) \beta_1(x_1)$$

with $\beta_t(x)$ the backward messages.




$$\begin{aligned} \beta_T(x_T) &= 1 \\ \beta_{t-1}(x_{t-1}) &= \sum_{x_t} q(x_t|x_{t-1}) \exp(R(x_t)) \beta_t(x_t) \end{aligned}$$

# KL control theory

The control computation is 'reduced' to a (graphical model) inference problem.

| | | |
|---|---|---|
| Dynamics: $p^t_{xy}(\pi)$ | $\rightarrow$ DP $\rightarrow$ | Bellman Equation |
| Cost: $C(\pi^{0:T}) = \langle R \rangle$ | | |
| $\downarrow$ | | $\downarrow$ |
| restricted class | | approximate $J$ |
| $\downarrow$ | | $\downarrow$ |
| Free dynamics: $q^t_{xy}$ | $\rightarrow$ approx inference $\rightarrow$ | Optimal $\pi$ |
| $C = KL(p\|\|q\exp(R))$ | | |

Optimal solution:

$$p(x^{1:T}|x^0) \quad = \quad \frac{1}{Z} q(x^{1:T}|x^0) \exp(R(x^{0:T}))$$

Intractable, but standard approximate inference problem.

# Link to contiuous path integral formulation

The previous continuous path integral control can be obtained as a special case of the KL control formulation.

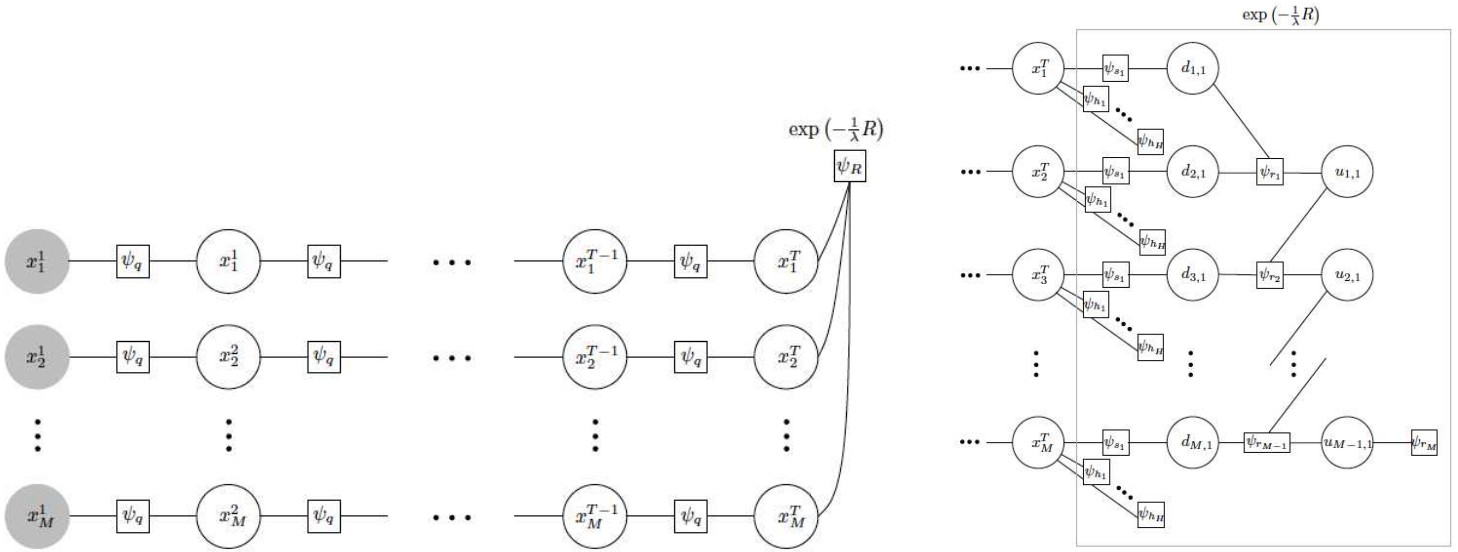
$$
\begin{aligned}
p(x_{t+dt}|x_t, u_t) &= \mathcal{N}(x_{t+dt}|x_t + f(x_t, t)dt + u_t dt, \nu) \\
q(x_{t+dt}|x_t) &= \mathcal{N}(x_{t+dt}|x_t + f(x, t)dt, \nu) \\
C(p|x_0) &= KL(p|q) - \langle R \rangle = \sum_{x^{dt:T}} p(x^{dt:T}|x^0) \left( \sum_{t=dt}^{T} \frac{1}{2} u_t^T \nu^{-1} u_t - R(x_t) \right)
\end{aligned}
$$

# Multi Agent cooperative game

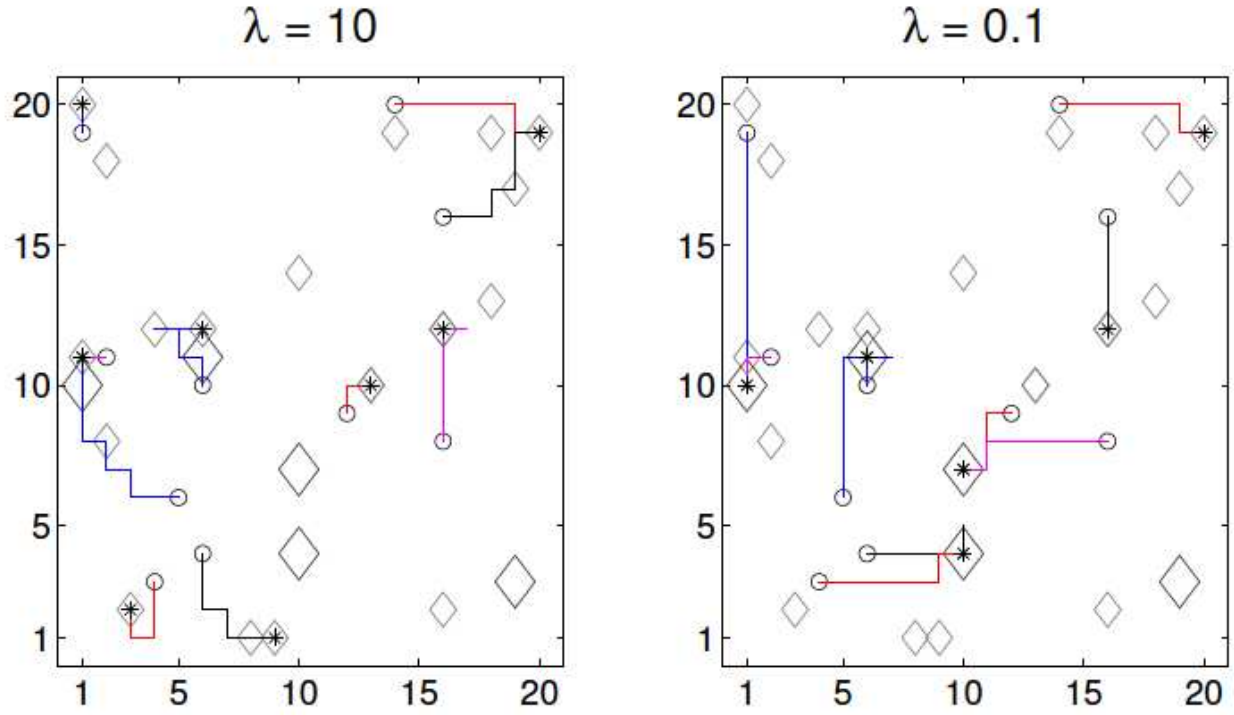
Model of cooperation: either hunt a hare alone or a stag together.

| | Stag | Hare |
|---|---|---|
| Stag | $\mathbf{3}, \mathbf{3}$ | $0, 1$ |
| Hare | $1, 0$ | $\mathbf{1}, \mathbf{1}$ |

We define the KL-stag-hunt game as a multi-agent version where agents move on a grid to hunt stag or hare.



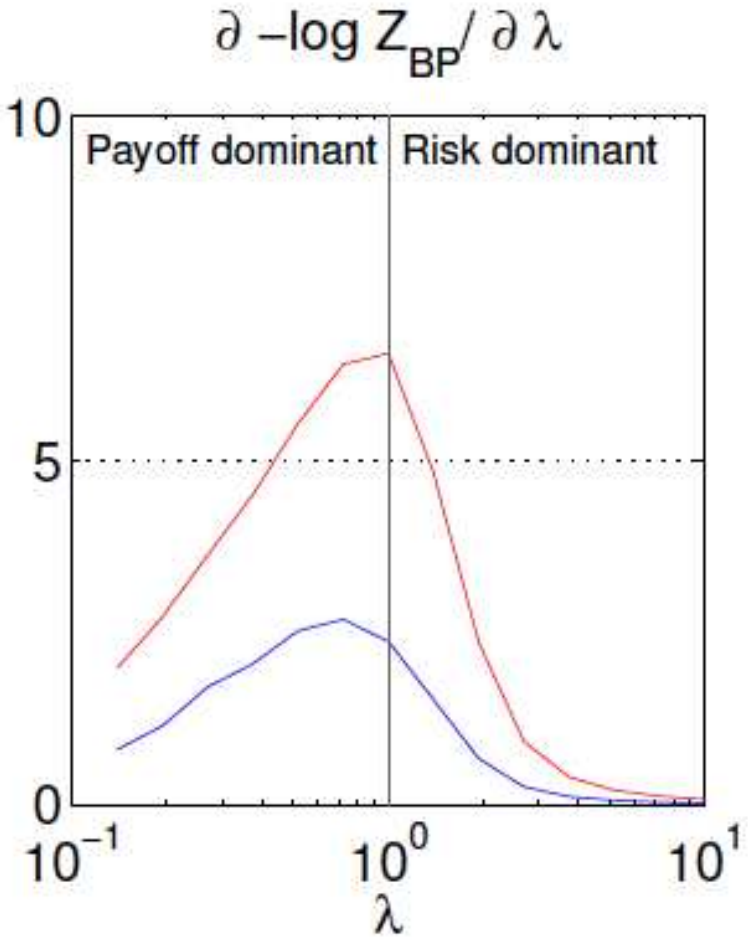
# Approximate inference of the KL-stag-hunt problem

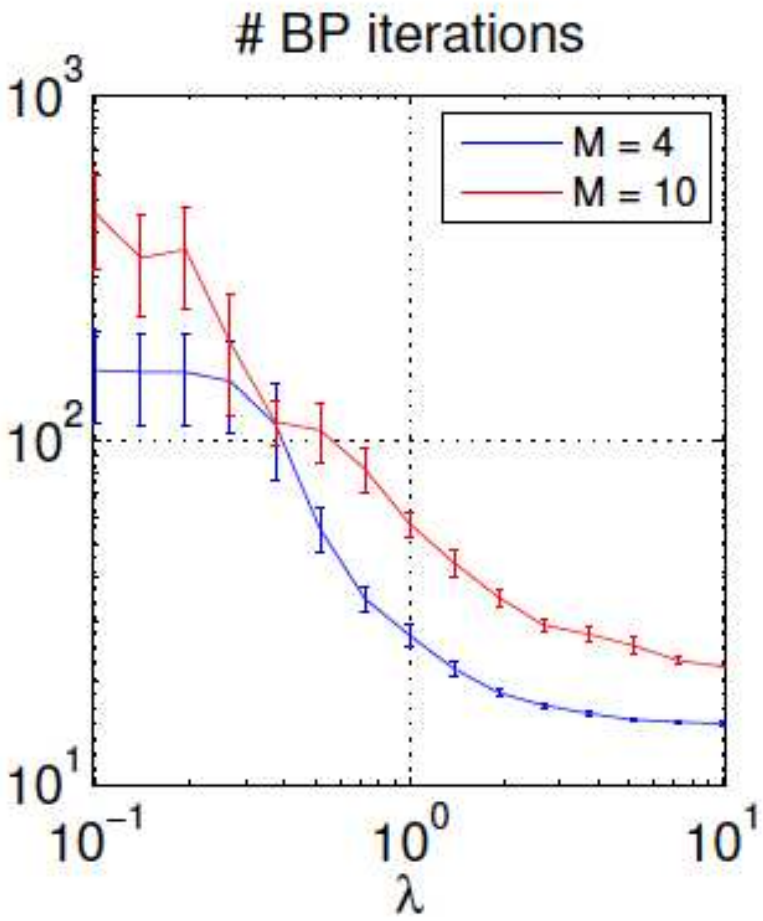



$M = 10$ agents, $N = 400$ locations, $10^{26}$ states per time slice

Sequential BP. If converges, converges in less than 500 iterations. Trajectories are marginal beliefs.

# Phase transition (?)



# Average cost KL control (Todorov 2006)

When $T \to \infty$ and $q$ ergodic the backward message recursion

$$\beta_{t-1}(x_{t-1}) = \sum_{x_t} H(x_{t-1}, x_t)\beta_t(x_t) \qquad H(x,y) = q(y|x)\exp(R(y))$$

becomes the computation of the Perron-Frobenius eigen pair $(\beta(\cdot), \lambda)$:

$$H\beta = \lambda\beta \qquad H(x,y) = q(y|x)\exp(R(x))$$

The optimal control satisfies

$$\begin{aligned} p(y|x) &= q(y|x)\exp(R(x))\frac{\beta(y)}{\lambda\beta(x)} \\ C &= -\log\lambda \\ J(x) &= -\log\beta(x) \end{aligned}$$

# KL-learning [Bierkens, Kappen 2012]

- Goal: find Perron-Frobenius solution $Hz = \lambda z$, with $H = [q(y|x)\exp(-R(x))]$, while stepping through state space according to $q$ and observing incurred cost.

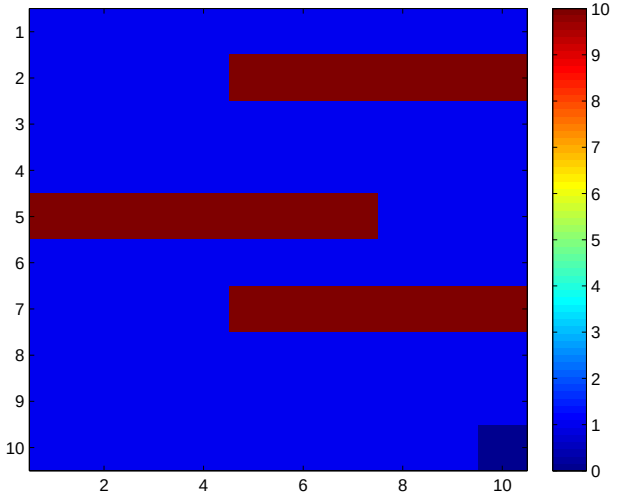
- Algorithm (KL-learning):

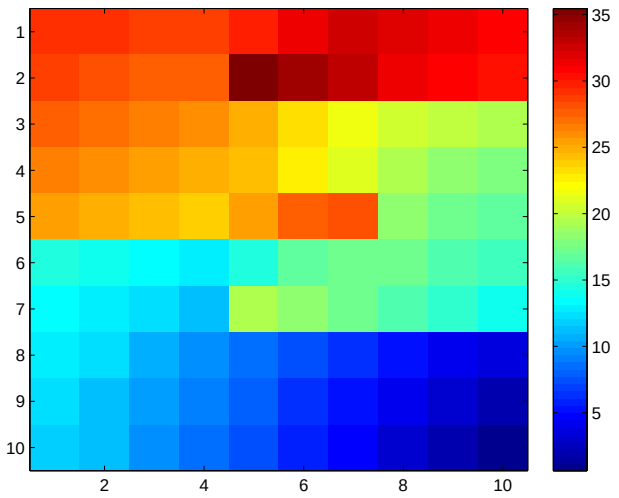
  $z \leftarrow (1/n, \ldots, 1/n)$, $\lambda > 0$, $x \leftarrow$ any state
  **for** $m = 1 : M$ **do**
    $y \leftarrow$ independent draw from $q(\cdot|x)$
    $\Delta \leftarrow \exp(-R(x))z(y)/\lambda - z(x)$
    $z(x) \leftarrow z(x) + \gamma\Delta$
    $\lambda \leftarrow \lambda + \gamma\Delta$
    $x \leftarrow y$
  **end for**

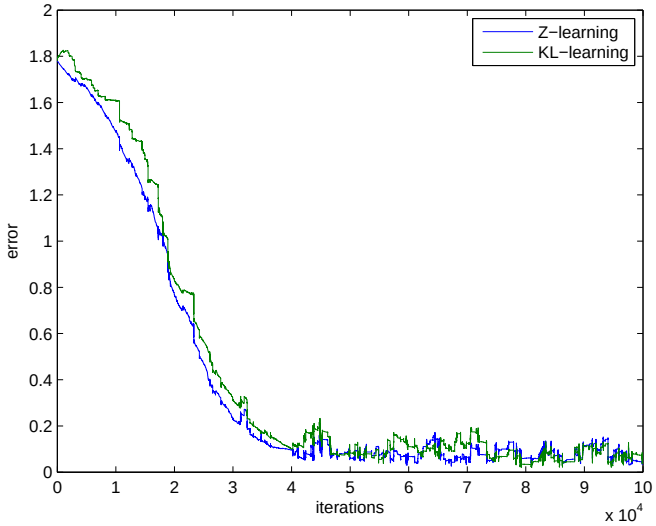
- Invariants: $z > 0$, $\lambda = ||z||_1$.

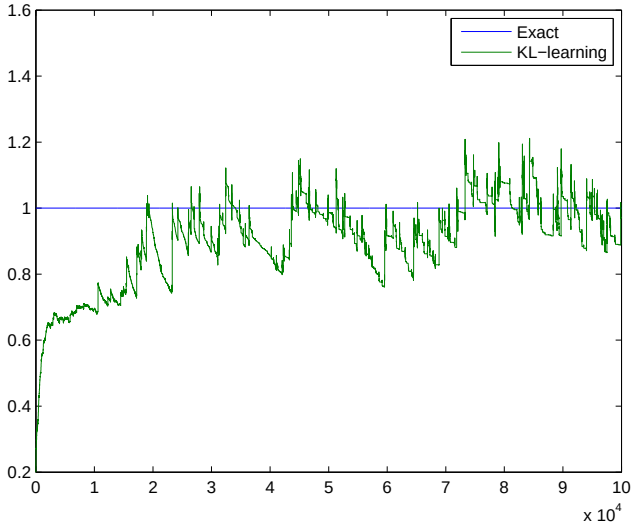
Generalization of $z$-learning (Todorov) to $\lambda \neq 1$

# Numerical experiment





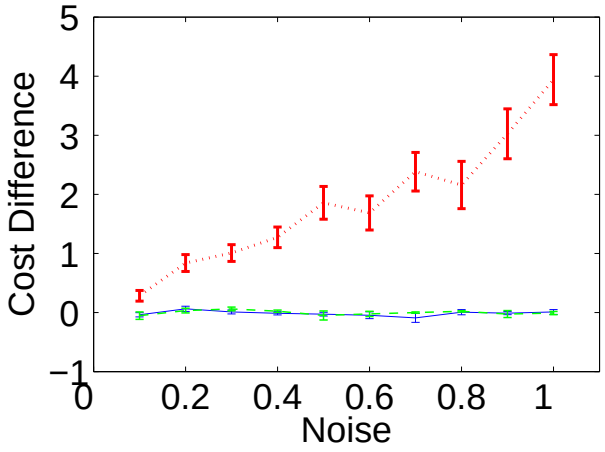


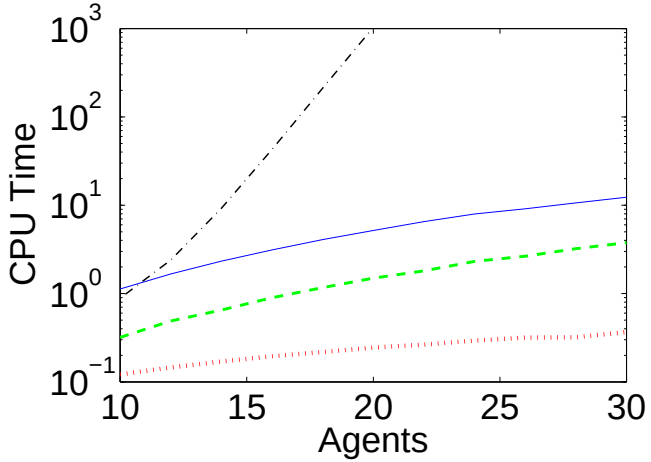



# Summary and discussion

Control as inference links control to machine learning and statistical physics

- efficient computational methods

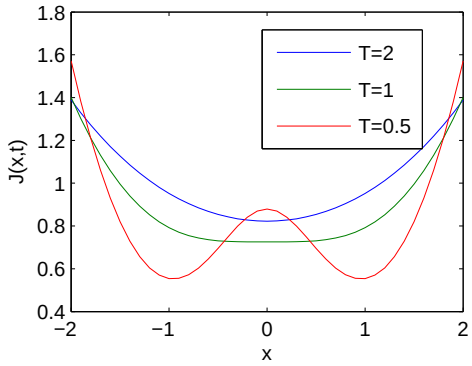


# Summary and discussion

Control as inference links control to machine learning and statistical physics

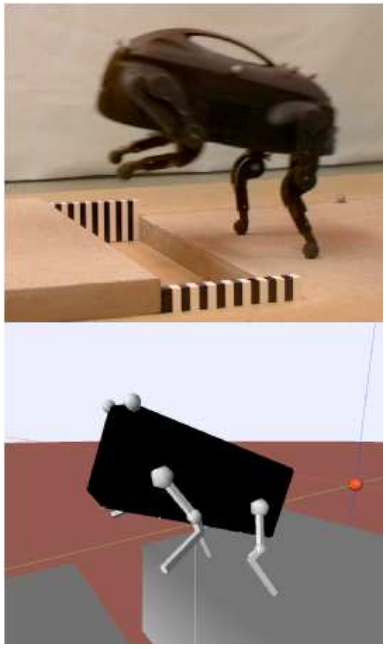
- efficient computational methods
- insight in the role of noise: phase transitions (delayed choice and collaboration)



# Summary and discussion

Control as inference links control to machine learning and statistical physics

- efficient computational methods
- insight in the role of noise: phase transitions (delayed choice and collaboration)
- favorable comparison with state-of-the-art RL methods in robotics (Theodorou 2010-2012)



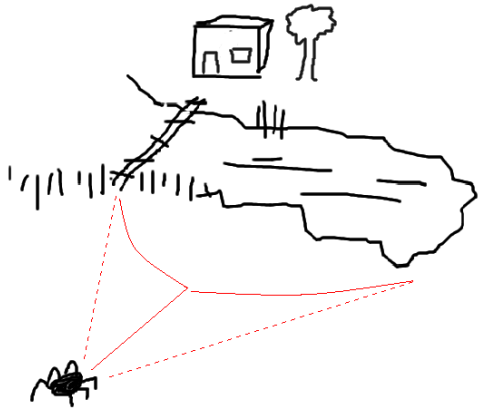
# Summary and discussion

Control as inference links control to machine learning and statistical physics

- efficient computational methods
- insight in the role of noise: phase transitions (delayed choice and collaboration)
- favorable comparison with state-of-the-art RL methods in robotics (Theodorou 2010-2012)

# Further reading

http://www.snn.ru.nl/~bertk/
http://www.snn.ru.nl



# References


[1] J Yong and X.Y. Zhou. *Stochastic controls. Hamiltonian Systems and HJB Equations*. Springer, 1999.

[2] J.J. Florentin. Optimal, probing, adaptive control of a simple Bayesian system. *International Journal of Electronics*, 13:165–177, 1962.

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

[4] E.J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University, 1971.

# Statistical mechanics of control theory

stochastic control theory
statistical physics
non-equilibrium systems
machine learning
transport theory, control of diffusions
quantum control
large deviation theory
neuroscience, robotics

Granada, 12-16 September 2012
`www.snn.ru.nl/cyberstat_granada`

Ari Arapostathis (U Texas)
Bill Bialek (Princeton)
Roger Brockett (Harvard)
Jena-Charles Delvenne (Louvain)
Volodya Chernyak (Detroit)
Karl Friston (UC London)
Francesco Guerra (Rome)
Ramon van Handel (Princeton)
Claudio Landim (Rouen)
Seth Lloyd (MIT)
Marc Mezard (Orsay)
Sanjoy Mitter (MIT)
Jun Morimoto (ART Japan)
Pablo Parrilo (MIT)
Juan Parrondo (Madrid)
Henrik Sandberg (Lund)
Devavrat Shah (MIT)
Naftali Tishby (Hebrew U)
Evangelos Theodorou (Washington)
Emanuel Todorov (Washington)

**Organizing committee:**

Misha Chertkov (Los Alamos)

Bert Kappen (Nijmegen)

Frank Redig (Delft)

Riccardo Zecchina (Torino)

# Other topics

- Variational approximation, $n$ joint arm (Kappen tutorial 2011)
- Sampling approach to control of robotics arm (van den Broek 2011)
- Coordination of continuous agents using MF and BP (Wiegerinck et al. 2006, van den Broek et al. 2006)
- Risk sensitive path integral control (van den Broek 2010)
- Inference and control (Kappen tutorial 2011)

# The variational method

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

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

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 variational method

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.

# The variational method

Because $V = 0, f = 0, g = 1$, the solution to uncontrolled dynamics is Gaussian [6]

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

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

$$u_i \quad = \quad \frac{1}{T-t} \left( \langle \theta_i \rangle - \theta_i^0 \right) \tag{7}$$

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

$$p(\vec{\theta}) = \frac{1}{\psi(\vec{\theta^0}, t)} \exp\left( -\sum_{i=1}^{n} (\theta_i - \theta_i^0)^2 / 2\nu(T-t) - \phi(\vec{\theta})/\nu \right) \tag{8}$$

[6] This is not exactly correct because $\theta$ 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 $2\pi$.

# The variational method

We compute the expectations $\left\langle \vec{\theta} \right\rangle$ by introducing a factorized Gaussian variational distribution $q(\vec{\theta}) = \prod_{i=1}^{n} \mathcal{N}((\theta_i|\mu_i, \sigma_i)$. We compute $\mu_i$ and $\sigma_i$ by by minimizing the KL divergence between $q(\vec{\theta})$ and $p(\vec{\theta})$:

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

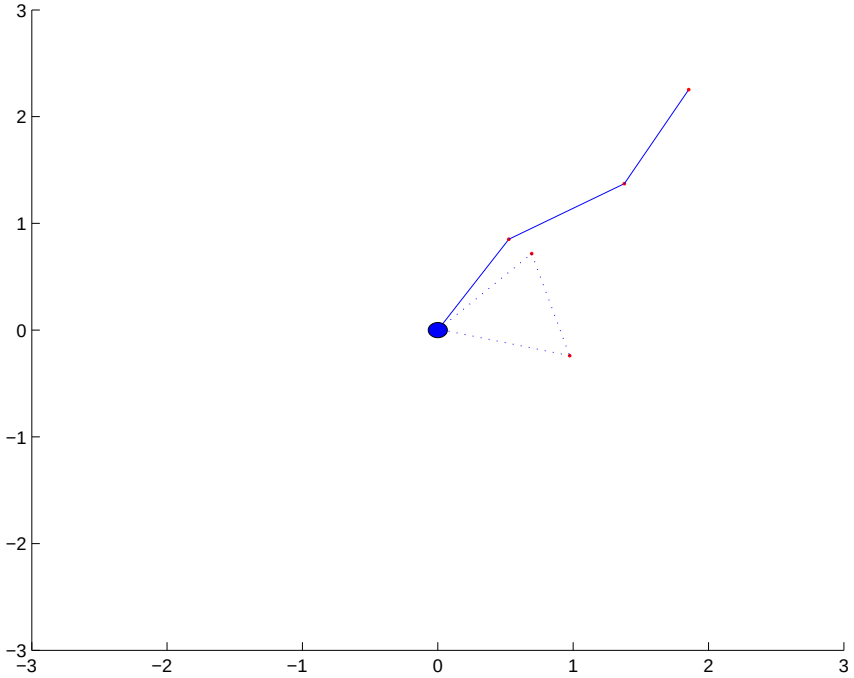
$$= -\sum_{i=1}^{n} \log \sqrt{2\pi\sigma_i^2} + \log \psi(\vec{\theta^0}, t) + \frac{1}{2\nu(T-t)} \sum_{i=1}^{n} \left( \sigma_i^2 + (\mu_i - \theta_i^0)^2 \right) + \frac{1}{\nu} \left\langle \phi(\vec{\theta}) \right\rangle_q$$

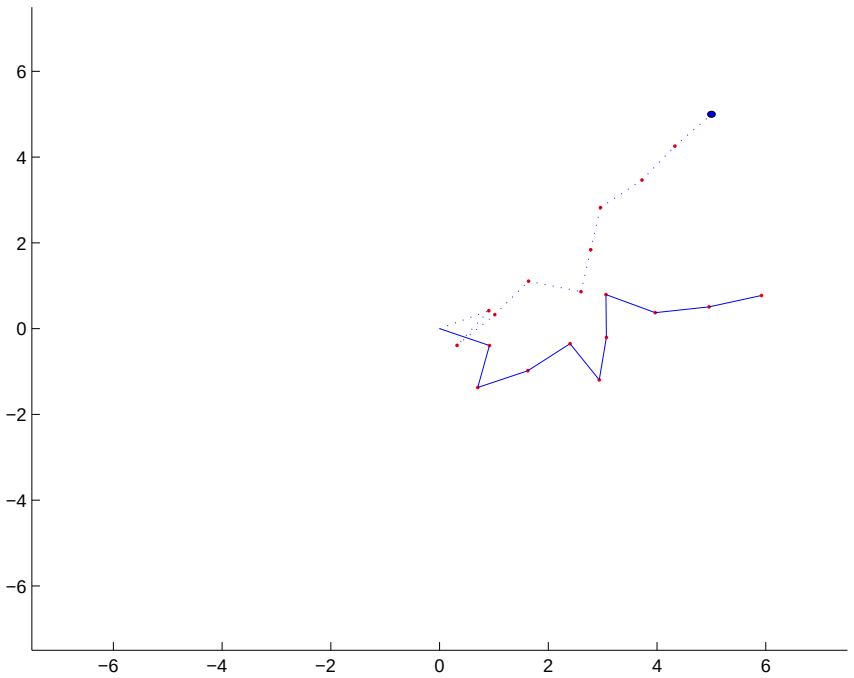
where we omit irrelevant constants. $\left\langle \phi(\vec{\theta}) \right\rangle$ can be computed in closed form. Setting the derivative of the KL with respect to $\mu_i$ and $\sigma_i^2$ equal to zero:

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

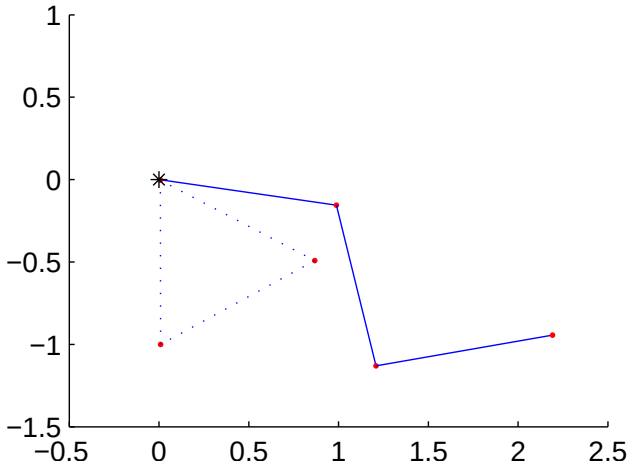
$$\frac{1}{\sigma_i^2} \leftarrow \frac{1}{\nu} \left( \frac{1}{(T-t)} + \alpha e^{-\sigma_i^2} - \alpha \left( \langle x_n \rangle - x_{\text{target}} \right) \cos \mu_i e^{-\sigma_i^2/2} - \alpha \left( \langle y_n \rangle - y_{\text{target}} \right) \sin \mu \right.$$

# The variational method

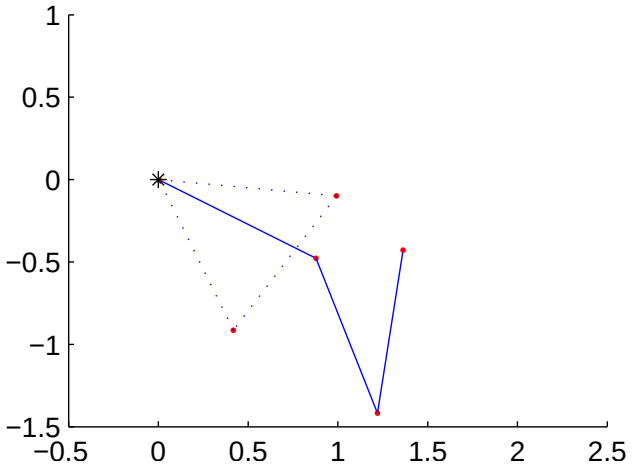


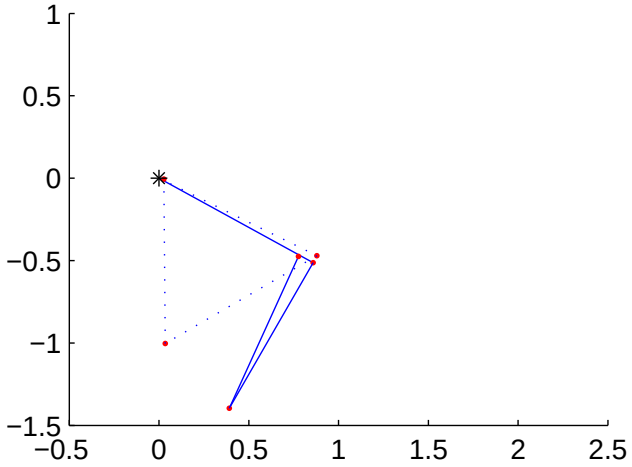
# The variational method



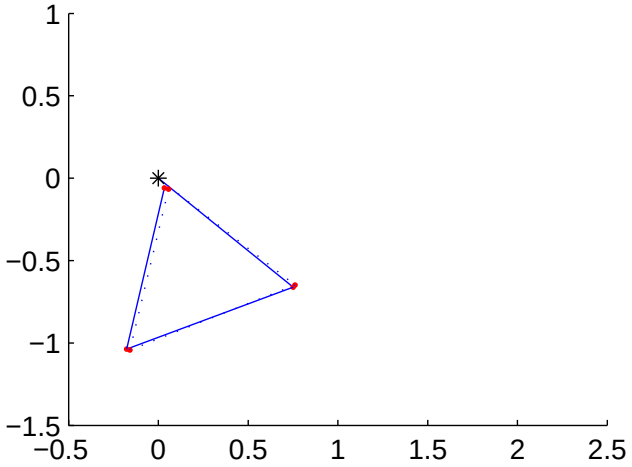

(a) $t = 0.05$




(b) $t = 0.55$

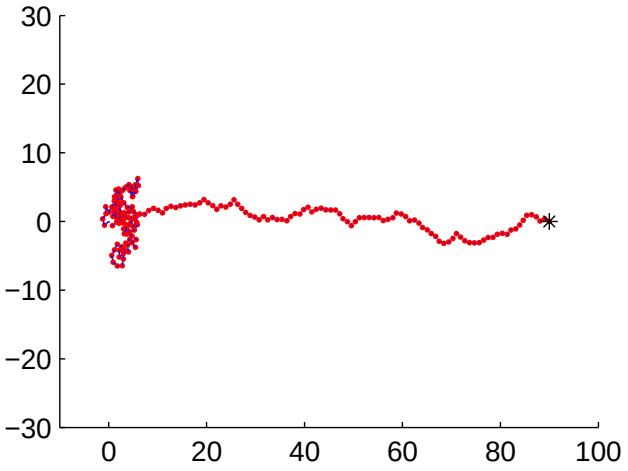



(c) $t = 1.8$

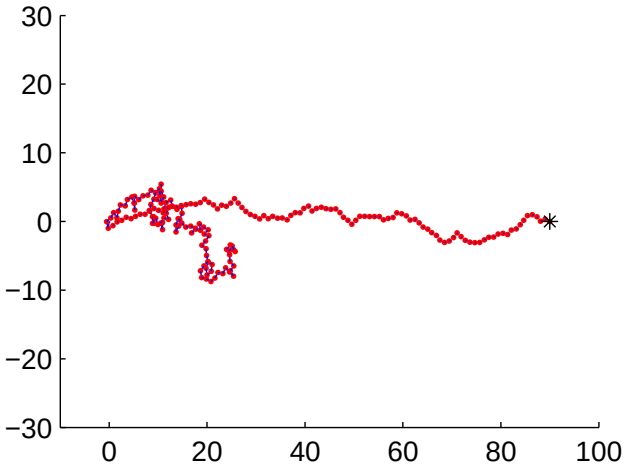



(d) $t = 2.0$

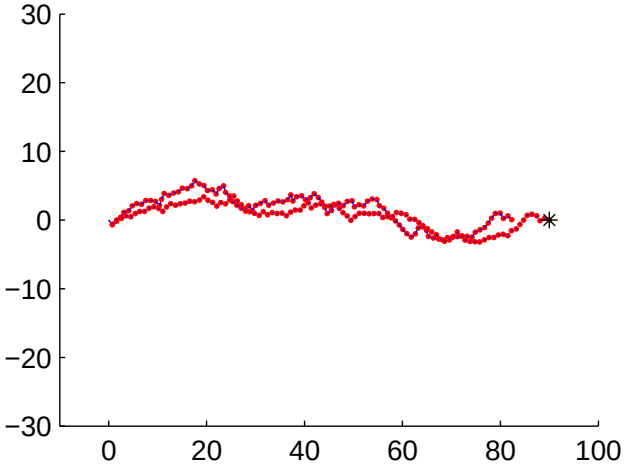
# The variational method



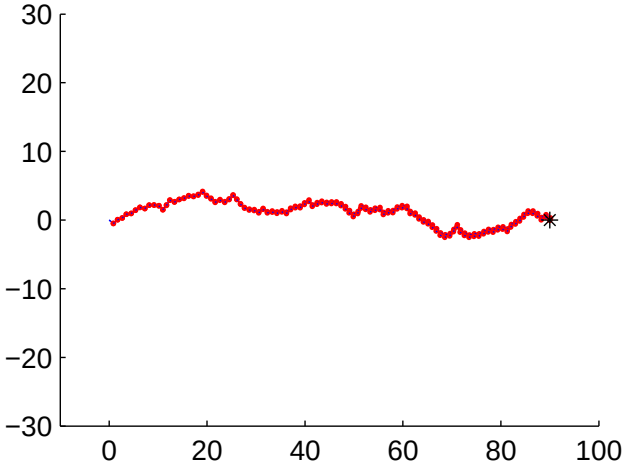

(e) $t = 0.05$




(f) $t = 0.55$




(g) $t = 1.8$




(h) $t = 2.0$

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 $\theta_i$ is steered independently to its target value $\langle \theta_i \rangle$ using the control law Eq. 7. 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.

# Coordination of agents

$n$ agents with independent dynamics

$$dx_\alpha = (f_\alpha(x_\alpha, t) + u_\alpha) + d\xi_\alpha, \quad \alpha = 1, \ldots, n$$

should coordinate their actions to minimize a cost at a future time $t = T$:

$$\phi(y_1, \ldots, y_n) \quad y_\alpha \in \{z_1, \ldots z_k\}$$

and $\phi = \infty$ elsewhere.

# Coordination of agents

Then,

$$\begin{aligned}
\Psi(x_1, \ldots, x_n, t) &= \int dy_1 \ldots dy_n \prod_\alpha \rho(y_\alpha, T|x_\alpha, t) \exp(-\phi(y_1, \ldots, y_n)/\nu) \\
&= \sum_{\vec{y}} \exp(-E(\vec{y}|\vec{x}, t)/\nu) \\
p(\vec{y}) &= \frac{1}{Z} \exp(-E(\vec{y}|\vec{x}, t)/\nu) \\
u_\alpha(\vec{x}, t) &= -\partial_{x_\alpha} J = \left\langle \frac{\partial \log \rho(y_\alpha, T|x_\alpha, t)}{\partial x_\alpha} \right\rangle
\end{aligned}$$

with $\vec{x} = (x_1, \ldots, x_n), \vec{y} = (y_1, \ldots, y_n)$.

$E$ has a graphical model structure if $\phi$ has.

# Pseudo code

Loop:

1. Compute the cost and its log derivative for each agent to move to each target:

$$\rho(z_i, T|x_\alpha, t), \quad i = 1, \ldots, k, \quad \alpha = 1, \ldots, n$$

   This path integral can be estimated using MC sampling or variational approximation.

2. Compute $u_\alpha$ using graphical model inference in $p(\vec{y})$ (exact, BP, MF).

# A simple 1d example

Intrinsic dynamics $f_\alpha = 0$, $V(x_1, \ldots, x_n) = 0$:

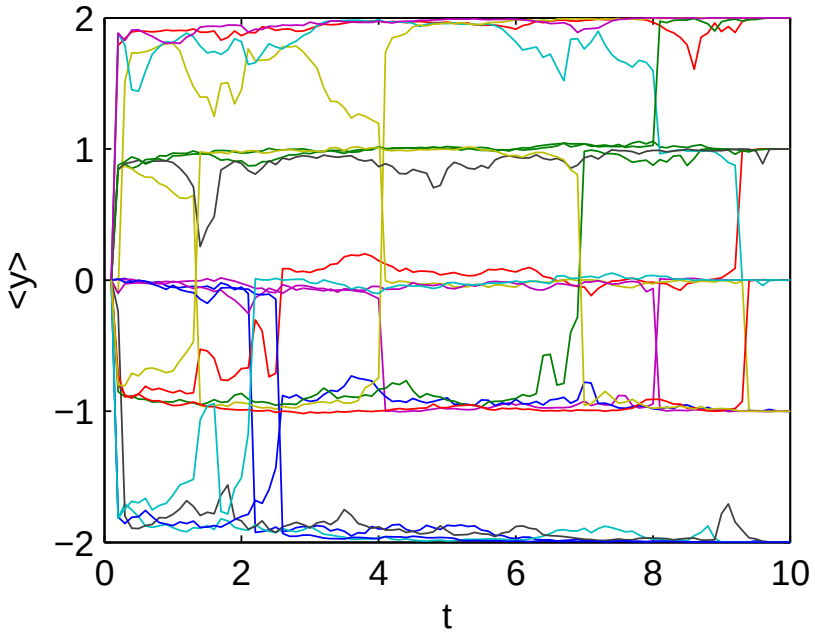
$$p(y_\alpha, T|x_\alpha, t) \propto \exp(-(y_\alpha - x_\alpha)^2/2\nu(T-t))$$

End cost $\phi(y_1, \ldots, y_n) = \sum_{j=1}^{k}(n_j(\vec{y}) - n_j)^2$, with $n_j(\vec{y})$ the # of agents that go to target $j$.

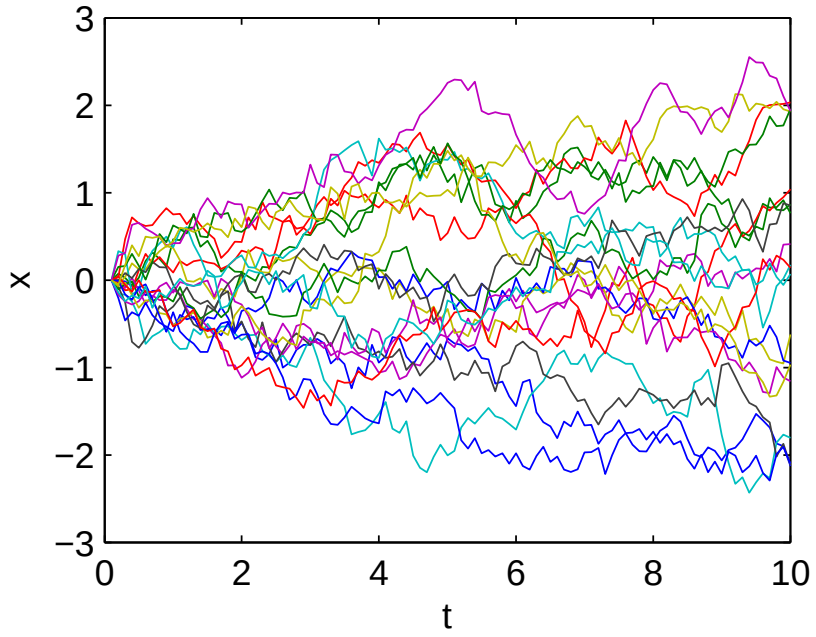
Optimal control is for agent $\alpha$ is

$$u_\alpha = \frac{1}{T-t}(\langle y_\alpha \rangle - x_\alpha)$$

# A simple 1d example

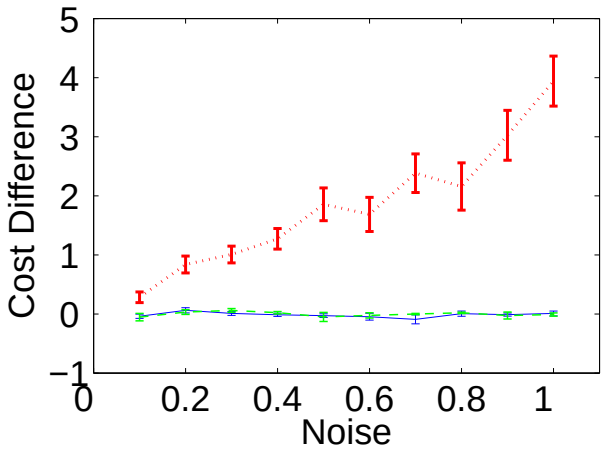


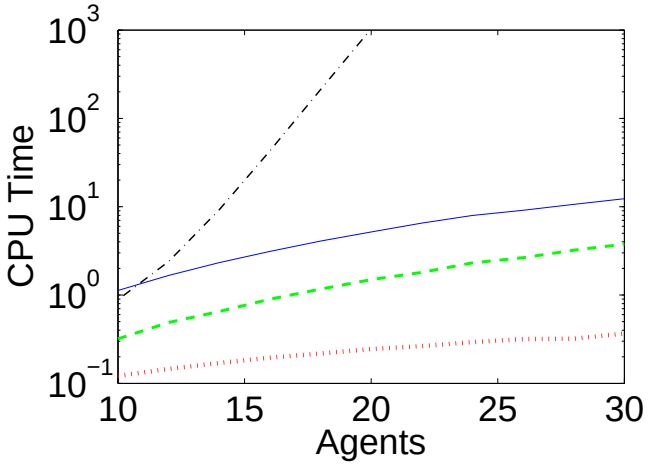

(i) Agent predicted target $\langle y_\alpha \rangle$




(j) Agent position $x$

# A simple 1d example




**Control cost**
greedy control (red)
MF control (blue)
BP control (green)

**CPU time**
exact control (black)
MF control (blue)
BP control (green)
greedy control (red)

# Nonlinear Coordination

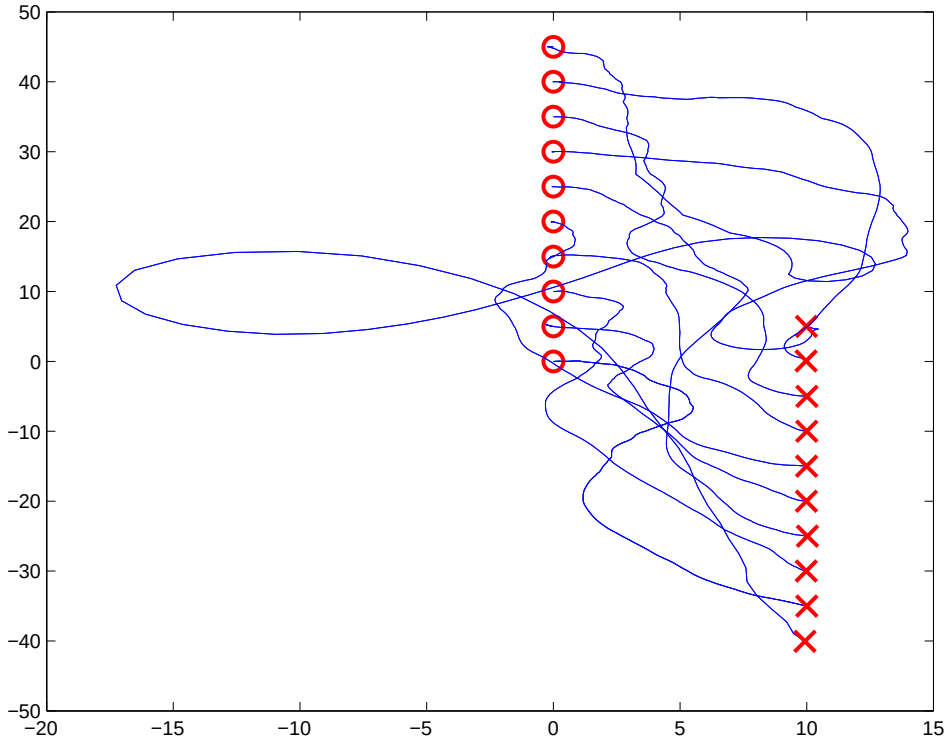
Agents $a = 1, \ldots, n$ in $2D$:

$$\begin{aligned} dx_a(t) &= v_a(t) \cos \varphi_a(t)\, dt \\ dy_a(t) &= v_a(t) \sin \varphi_a(t)\, dt \\ dv_a(t) &= u_a(t)dt + d\xi_a(t) \\ d\varphi_a(t) &= \omega_a(t)dt + d\zeta_a(t) \end{aligned}$$

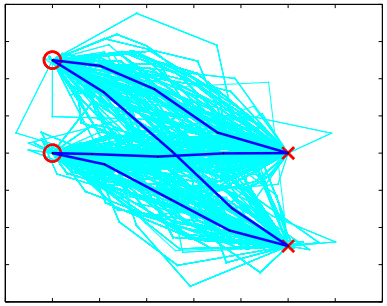
Initial states O, $v_a(0) = 0$, $\varphi_a(0) = 0$
Targets X, $v_a(T) = 0$, $\varphi_a(T) = 0$

Sample paths specified at
$t_i = t + i\, dt$,
$i = 0, \ldots, 6$, $dt = (T - t)/6$

Example of 10 agents & 10 targets:




Sample paths:



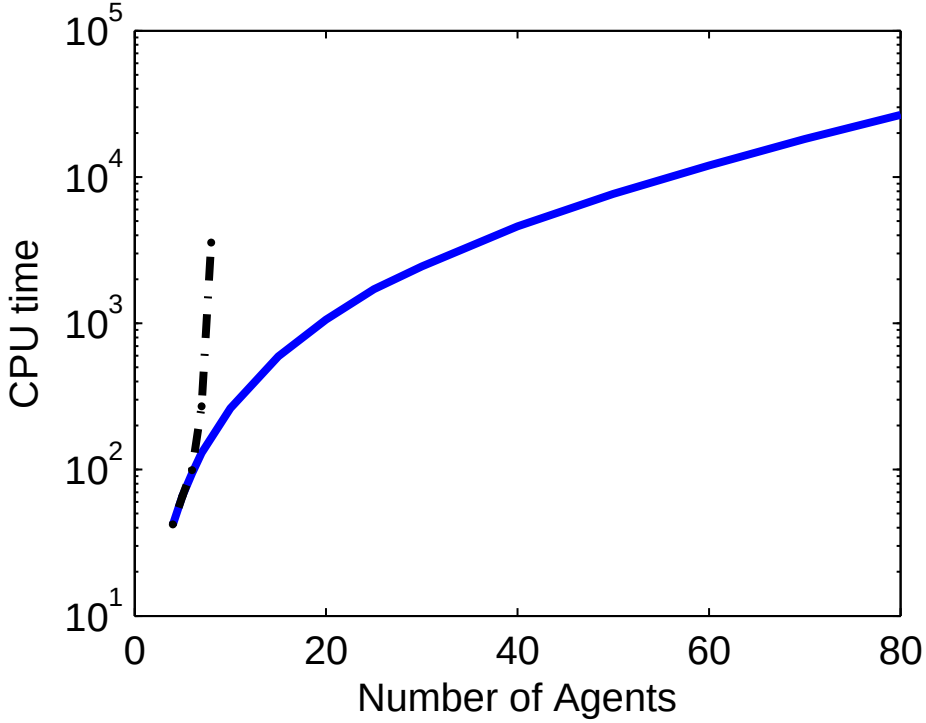
# Computation Time

Inference methods:

Junction Tree ($\cdot - \cdot$)
MF (—)

(100 sample paths per agent-target)

CPU time (s) vs. number of agents:




(# agents = # targets)

JT : exponential in number of agents
(intractable for # agents $> 10$)
MF : polynomial in number of agents

# Risk sensitive control

It is relatively straightforward to generalize the path integral method to optimize a cost of the form

$$\begin{aligned} \tilde{C} &= \phi(x_T) + \int \frac{1}{2} u^T R u + V(x) \\ C &= \frac{1}{\theta} \log \left\langle \exp(\theta \tilde{C}) \right\rangle \end{aligned}$$

For $\theta = 0$ the risk neutral control is recovered. For $\theta$ small:

$$C = \left\langle \tilde{C} \right\rangle + \frac{\theta}{2} \left( \left\langle \tilde{C}^2 \right\rangle - \left\langle \tilde{C} \right\rangle^2 \right) + h.o.$$

$\theta > 0$ is risk averse, $\theta < 0$ is risk seeking.

vd Broek et al. UAI 2010

# Risk sensitive control

We illustrate the behavior for the (well known) LQ case. $V = f = 0, \phi = \alpha/2x^2$.

The optimal control is given by

$$u = \frac{-\alpha x}{R + \alpha(T-t)(1-\nu R\theta)}$$

For $\theta < 0$ control is weaker
For $0 < \theta < 1/R\nu$ control is stronger
In both cases control increases with time.

For $\theta > 1/R\nu$, control is only well-defined when the denominator is positive:

$$\alpha(T-t) < \frac{R}{\nu R\theta - 1}$$

Control decreases with time. For larger time-to-go, the expected cost is infinite.

vd Broek et al. UAI 2010

# Inference and control

As an example of the intricacies of joint inference and control , consider the simple LQ control problem [2, 3]

$$dx = \alpha u dt + d\xi \tag{9}$$

$$C(x_0, \theta_0, u(0 \to T)) = \left\langle \phi(x(T)) + \int_0^T dt R(x, u, t) \right\rangle \tag{10}$$

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

Although $\alpha$ is unobserved, we have a means to observe $\alpha$ 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 $\alpha$ using the Bayes formula:

$$p_{t+dt}(\alpha|dx, u) \propto p(dx|\alpha, u) p_t(\alpha) \tag{11}$$

$p(dx|\alpha, u)$ is Normal in $dx$ with variance $\nu dt$
$p_t(\alpha)$ our belief at time $t$ about the values of $\alpha$

The information that we receive about $\alpha$ increases with $u$, because the $\alpha u dt$ term dominates the $d\xi$ term. 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 $\theta_t$ (sufficient statistics) that describe $p_t(\alpha) = p(\alpha|\theta_t)$ and $\theta_0$ known, which describes our initial belief in the possible values of $\alpha$. The cost that must be minimized is

$$C(x_0, \theta_0, u(0 \to T)) = \left\langle \phi(x(T)) + \int_0^T dt R(x, u, t) \right\rangle \tag{12}$$

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

NB: the average over $\alpha$ depends on $\theta_t$ which is not known beforehand.

For simplicity, consider the example that $\alpha$ 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. 11 implies a dynamics for $\theta$:

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

7

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

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

with boundary condition $J(z,T) = \phi(x)$ (NB independent of $\theta$).

[7] The rhs of the Bayes rule is

$$p(dx|\alpha,u)p(\alpha|\theta_t) \propto \exp\left(-\frac{(dx - \alpha u dt)^2}{2\nu dt}\right) \exp(\alpha\theta_t) \propto \exp\left(\frac{dx\alpha u}{\nu} + \alpha\theta_t\right) = \exp\left(\alpha\left(\theta_t + \frac{dxu}{\nu}\right)\right)$$

The result is

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

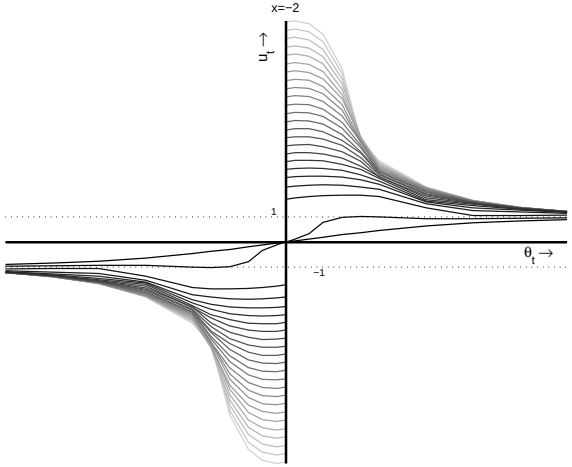
[8] with boundary conditions $J(x, \theta, T) = \phi(x)$.

Thus, the dual control problem (joint inference on $\alpha$ and control problem on $x$) has become an ordinary control problem in $x, \theta$ (Florentin, 1962).

Note that if $R, \phi$ are quadratic and $\alpha$ is known, this is an LQ problem. However, when $\alpha$ is not known, the corresponding dual control problem is not LQ (because of the additional $u$ dependent terms).

[8] The expectation values appearing in this equation are conditioned on $(x_t, \theta_t)$ and are averages over $p(\alpha|\theta_t)$ and the Gaussian noise. $\langle dx \rangle_{x,\theta} = \bar{\alpha} u dt$, $\langle d\theta \rangle_{x,\theta} = \frac{\bar{\alpha} u^2}{\nu} dt$, $\left\langle dx^2 \right\rangle_{x,\theta} = \nu dt$, $\left\langle d\theta^2 \right\rangle_{x,\theta} = \frac{u^2}{\nu} dt$, $\langle dx d\theta \rangle = u dt$, with $\bar{\alpha} = \tanh(\theta)$ the expected value of $\alpha$ for a given value $\theta$.

# Probing

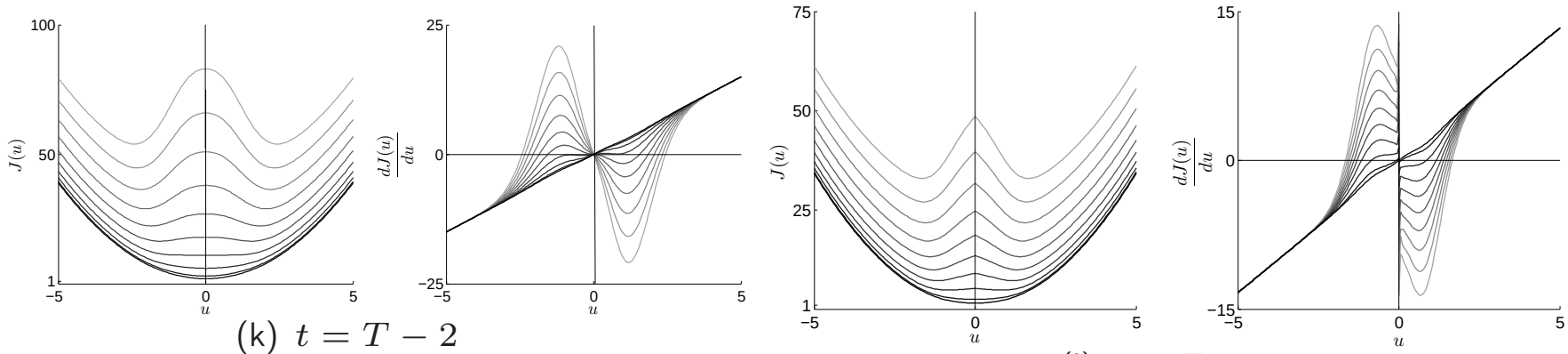



Dual control solution with end cost $\phi(x) = x^2$ and path cost $\int_t^{tf} dt' \frac{1}{2} u(t')^2$ and $\nu = 0.5$. Plot shows the deviation of the control from the certain case: $u_t(x, \theta)/u_t(x, \theta = \pm\infty)$ as a function of $\theta$ for different values of $t$ and $x = 2$. The curves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

'Probing': $u$ is much larger when $\alpha$ is uncertain ($\theta$ small) then when $\alpha$ is certain $\theta = \pm\infty$.

# Symmetry breaking and non-differentiability of $J$

The observed probing behavior arises as the result of a symmetry breaking in the right hand side of the Bellman equation.



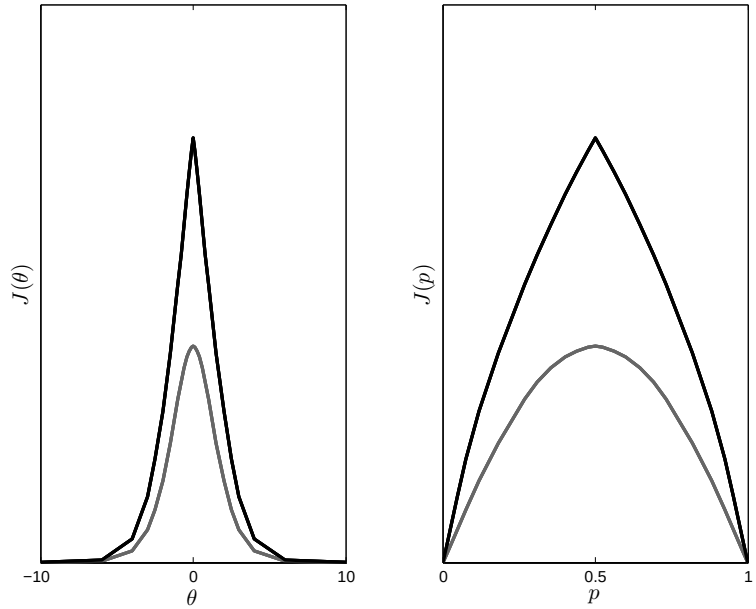

(k) $t = T - 2$ (l) $t = T - 3$

Figure 1: Rhs of the Bellman equation as a function of $u$ and its derivative for $\theta = 0$. The different curves correspond to different values of $x$. Explorative behavior ($u \neq 0$) arises in the no-knowledge state $\theta = 0$ by proposing non-zero controls. The singularity is absent at $t = T - 2$ and present starting from $t = T - 3$.

# Symmetry breaking and non-differentiability of $J$

As a result of the local minima in the Bellman optimization, the optimal value function is not differentiable.

The optimal cost-to-go is convex in the belief [4].




Left) $J_t(x, \theta)$ for $t = T - 2, x = -2$ (grey) and $t = T - 2, x = -6$ (black) versus $\theta$ Right) Same as left, but as a function of the belief $p = p(b = 1|\theta)$.