Optimal control as a graphical model inference problem

Bert Kappen
SNN Radboud University
Nijmegen

May 2010

Stochastic optimal control theory

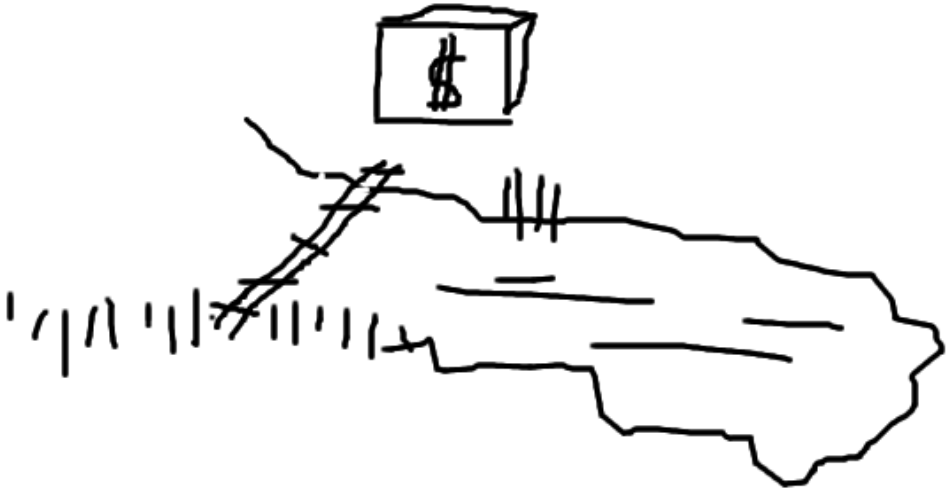
Control theory: how to act (now) to optimized future rewards

Idea: Use control theory to model intelligent (animal, human, robot, computer) behavior
- cooking together with your home robot
- anticipating human machine interface
- controlling when to learn

Noise and uncertainty plays dominant role:
- limited sensors
- at best probabilistic description of likely future events
- intractable

Tractable approaches are too simple:
- linear quadratic case
- deterministic case

Stochastic optimal control theory



Optimal solution is noise dependent

Outline

Introduction to control theory and review of path integral control

- delayed choice
- cooperating agents

Introduction to KL control theory

- path integrals as a special case
- opponent modeling, ficticious play, variational approximation
- stag hunt game

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.

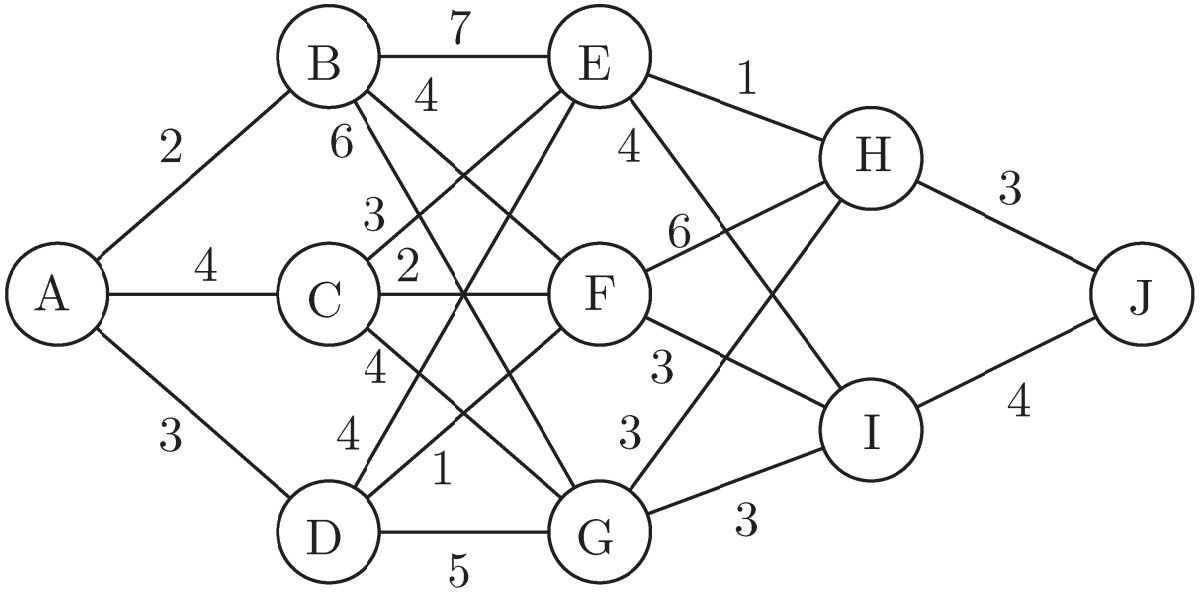
Given $x_{t=0} = x_0$ and $u_{0:T-1} = u_0, u_1, \ldots, u_T - 1$, we can compute $x_{1:T}$.

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. u is the choice of path and $R(u,t)$ is path cost

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

Discrete time control

One can recursively compute the solution by defining the optimal cost-to-go

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

This is called the *Bellman Equation*.

Computes $u(x, t)$ for all intermediate x, t backwards in time.

The optimal control at $t = 0$ is $u(x_0, 0)$.

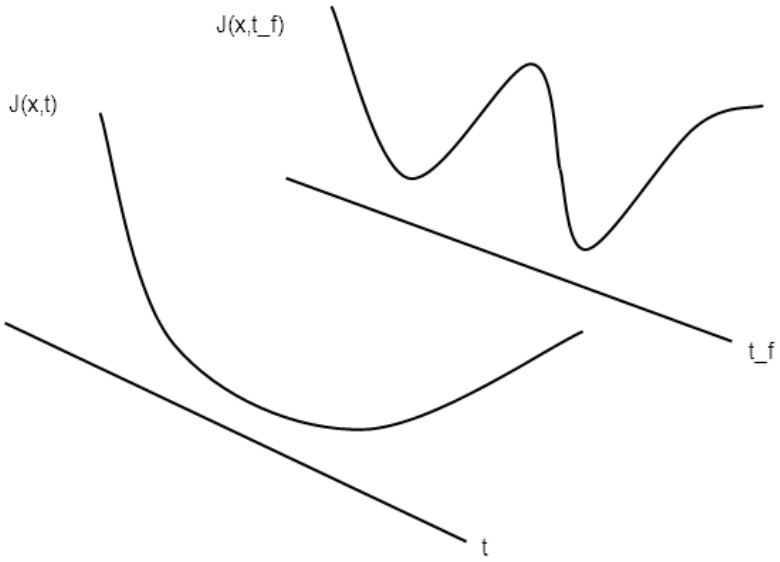
Continuous stochastic case

$$
\begin{aligned}
dx &= f(x,u,t)dt + d\xi, \qquad \left\langle d\xi d\xi^T \right\rangle = \nu dt \\
C &= \left\langle \phi(x(T)) + \int_0^T dt R(t,x(t),u(t)) \right\rangle
\end{aligned}
$$

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

solved backwards in time 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 $\phi(x)$.

The path integral solution to stochastic optimal control

Consider additive linear control and quadratic control cost:

$$\begin{aligned} dx &= (f(x,t)+u)dt + d\xi \\ R(x,u,t) &= V(x,t) + \frac{1}{2}u^T R u \end{aligned}$$

with with $\left\langle d\xi d\xi^T \right\rangle = \nu dt$ and $\lambda = R\nu$.

With $R = 1$:

$$\begin{aligned} -\partial_t J &= \min_u \left(\frac{1}{2}u^2 + V + (f+u)\partial_x J + \frac{1}{2}\nu\partial_x^2 J \right) \\ &= -\frac{1}{2}(\partial_x J)^2 + V + f\partial_x J + \frac{1}{2}\nu\partial_x^2 J \end{aligned}$$

with boundary condition $J(T,x) = \phi(x)$ and $u = -\partial_x J$.

The logarithmic transformation

$$J(x,t) = -\nu \log \Psi(x,t)$$

The HJB equation becomes linear in Ψ:

$$\partial_t \Psi = -H\Psi, \qquad H = -\frac{V}{\nu} + f\partial_x + \frac{1}{2}\nu\partial_x^2$$

with boundary condition $\Psi(T,x) = \exp(-\phi(x)/\nu)$.

Reversing time

Let $\rho(y,\tau|x,t)$ describe a diffusion process defined by the Fokker-Planck equation

$$\partial_\tau \rho = H^\dagger \rho$$

with $\rho(y,t|x,t) = \delta(y-x)$.

It can be shown that

$$\begin{aligned}
\Psi(x,t) &= \int dy \rho(y,T|x,t) \exp(-\phi(y)/\nu) \\
\rho(y,T|x,t) &= \int [dx]_x^y \exp\left(-S_{\rm path}/\nu\right) \\
J(x,t) &= -\nu \log \int [dx]_x \exp\left(-\frac{1}{\nu} S(x(t \to t_f))\right) \\
S(x(t \to t_f)) &= \phi(x(t_f)) + \int d\tau \frac{1}{2}(\dot{x}(\tau) - f(x(\tau),\tau))^2 + V(x(\tau),\tau)
\end{aligned}$$

where the path integral $\int [dx]_x$ is over all trajectories starting at x.

The path integral formulation

$$J(x,t) \quad = \quad -\nu \log \int [dx]_x \exp\left(-\frac{1}{\nu} S(x(t \to t_f))\right)$$

The path integral is a log partition sum and therefore can be interpreted as a free energy. S is the energy of a path and ν the temperature:
1) noise dependent solution (high and low temperature, phase transition)
2) use standard (stat mech) methods for approximate computation
- MC sampling
- variational, BP,...

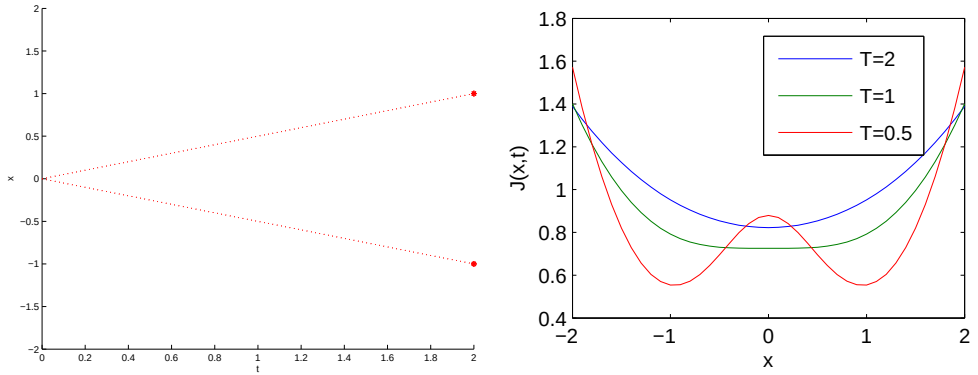
Delayed choice

$$x_{t+dt} = x_t + u_t dt + d\xi_t \qquad \left\langle \xi_t^2 \right\rangle = \nu dt$$

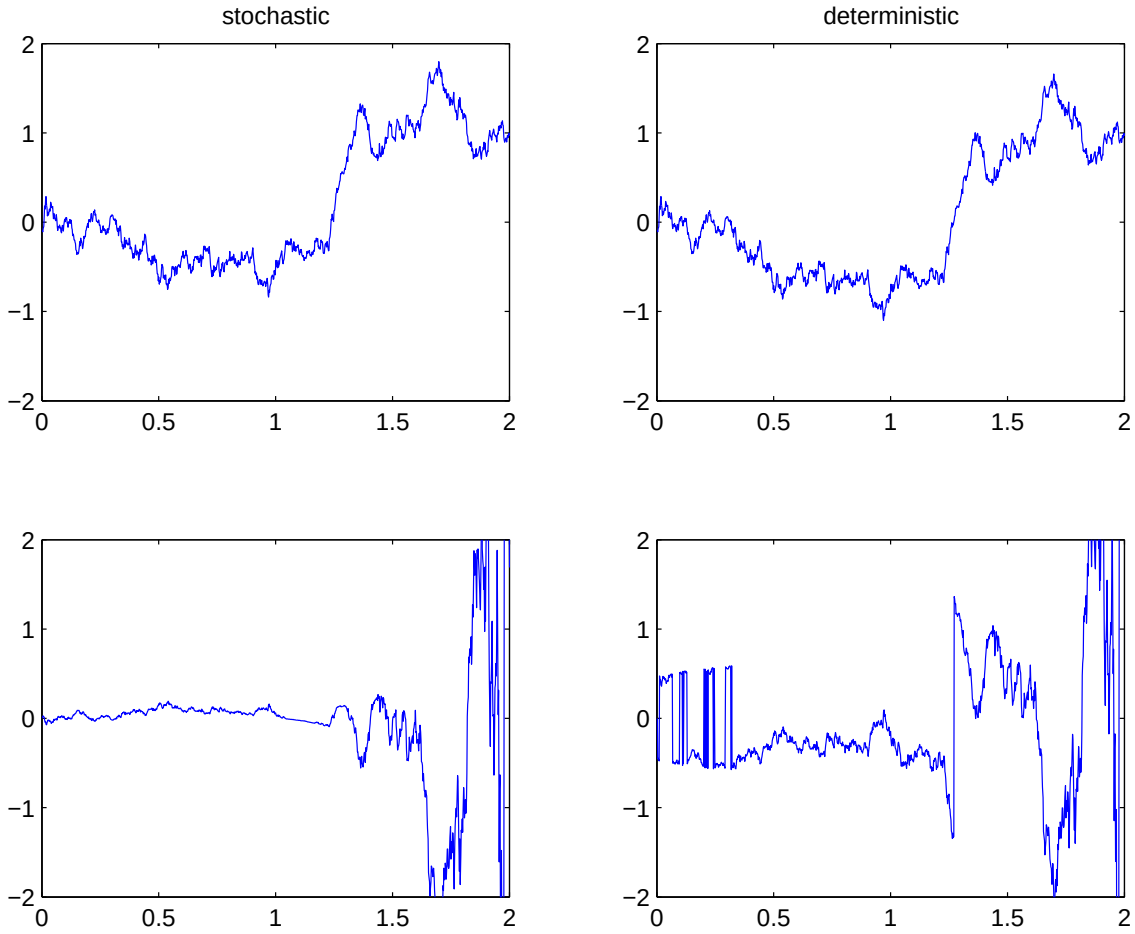
$V = 0$ and end cost ϕ encodes two targets at $t = T$.

$$J(x, t=0) = \frac{1}{T}\left(\frac{1}{2}x^2 - \nu T \log 2\cosh\frac{x}{\nu T}\right)$$

$$u(x) = \frac{1}{T}\left(\tanh\frac{x}{\nu T} - x\right)$$

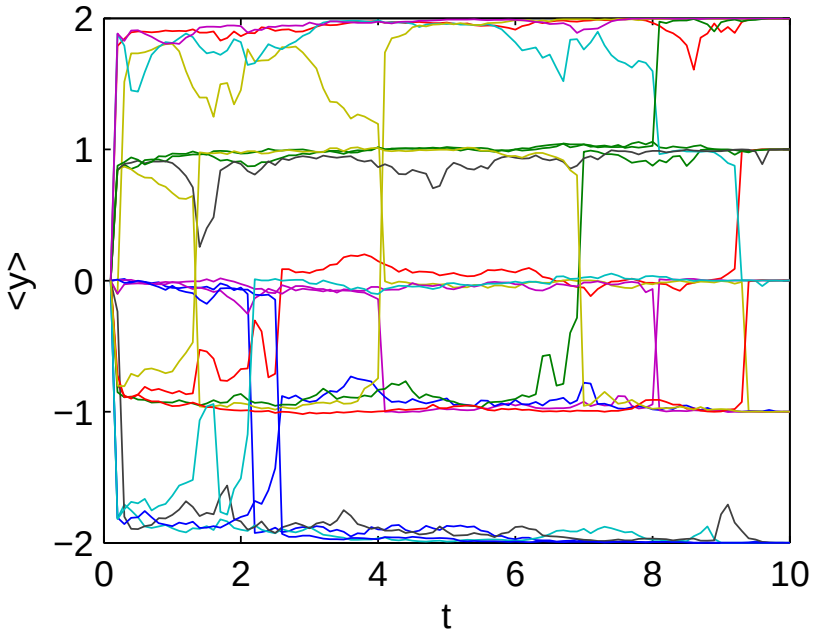


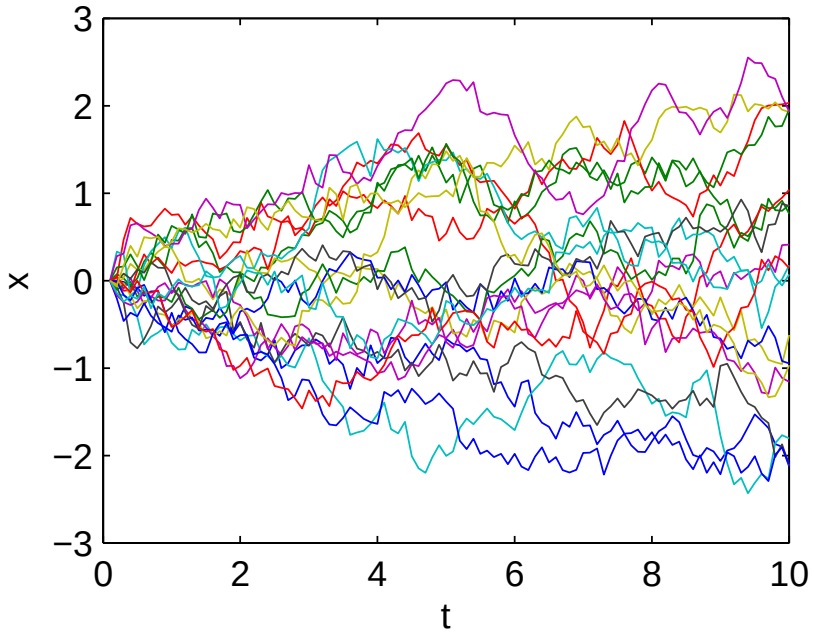
Delayed choice



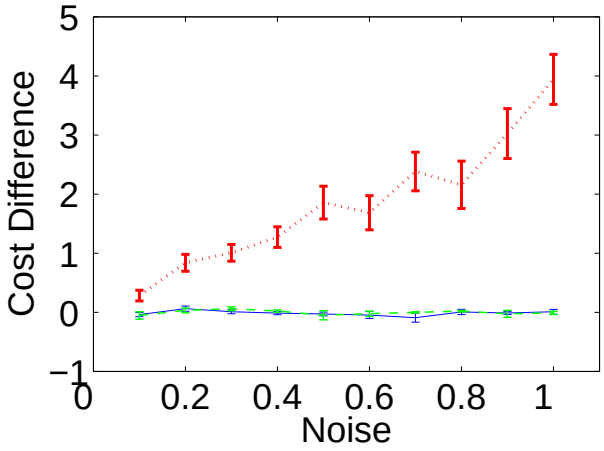
Prediction: when the future is uncertain, delay your decisions.

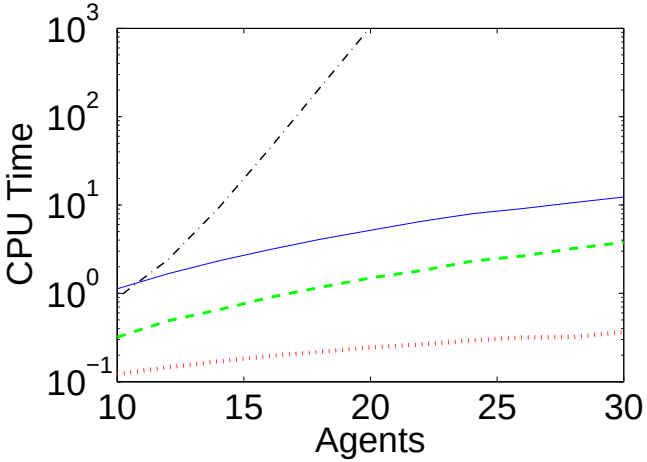
Firemen extinguishing fires





Firemen extinguishing fires





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)

KL control theory

A different view and a possible generalization.

Approximate inference

Write $p(x) = \frac{1}{Z}\exp(-E(x))$.

p is given by minimizing the

$$KL(p||\exp(-E)) = \sum_x p(x)\log\frac{p(x)}{\exp(-E(x))}$$

wrt p subject to normalization constraint.

Approximate $p(x) \Leftrightarrow$ approximate KL or restrict minimization
- variational
- BP, CVM

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

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

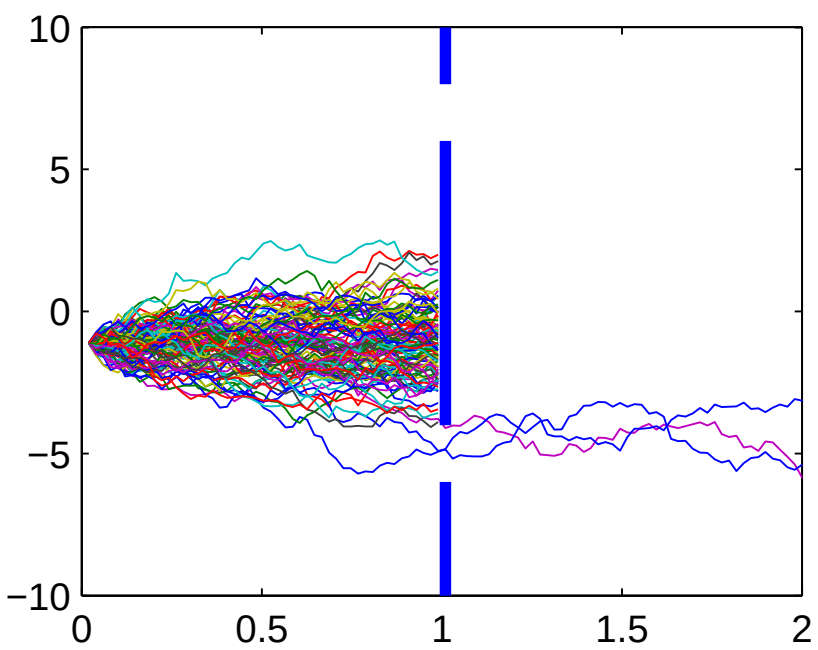
$q(x_{t+1}|x_t)$ implements the allowed moves.

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

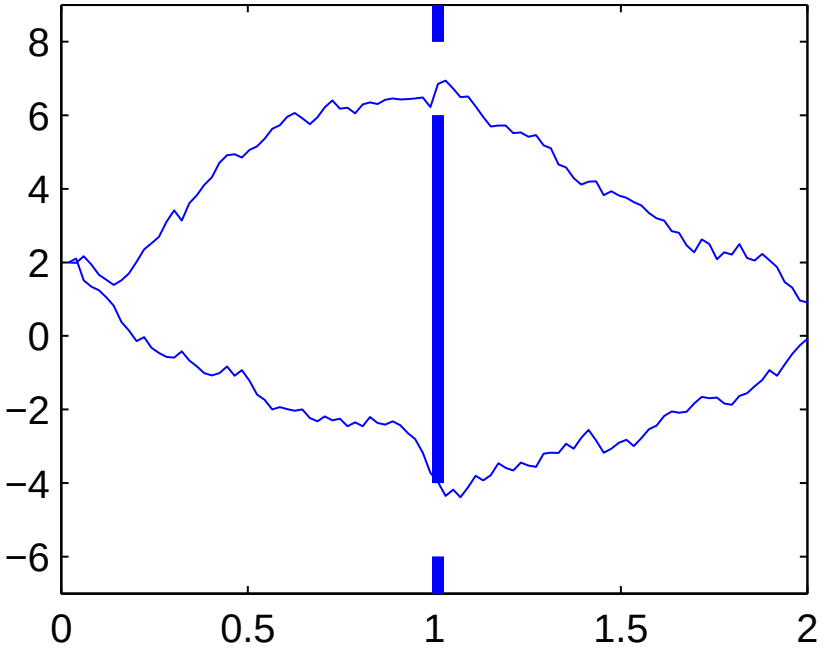
The KL control problem is to find the probability distribution $p(x_{1:T}|x_0)$ that minimizes

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

KL control theory



(a) Sample paths under q



(b) Sample paths under p

KL control theory

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

The optimal solution for p is found by minimizing C wrt p. The solution and the optimal control cost are

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

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

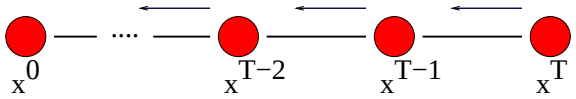
Intractable but standard inference problem.

KL control theory

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

$$p(x_1|x_0) = \sum_{x_{2:T}} p(x_{1:T}|x_0) \propto q(x_1|x_0) \exp(R(x_1)) \beta_1(x_1)$$

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



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

Continuous case

Consider again

$$x_{t+dt} = x_t + f(x_t, t)dt + u_t dt + d\xi_t$$

In terms of the KL control formulation

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

Agents: a distributed approach

In the case of agents, the uncontrolled dynamics q factorizes over the agents:

$$q(x^1_{1:T}, x^2_{1:T}, \ldots | x^1_0, x^2_0, \ldots) = q^1(x^1_{1:T}|x^1_0) q^2(x^2_{1:T}|x^2_0) \ldots$$

However, the reward R is a function of the states of all agents and can be different for each agent.

Opponent modeling: each agent assumes a model according to which the other agents behave.

$$\begin{aligned} C^1(p^1|x^1_0, x^2_0) &= KL(p^1||q^1) - \left\langle R^1 \right\rangle_{p^1, \hat{p}^2} \\ p^1(x^1_{1:T}|x^1_0, x^2_0) &= \frac{1}{Z^1(x_0)} q^1(x^1_{1:T}|x^1_0) \exp\left(\left\langle R^1 \right\rangle_{\hat{p}^2} \right) \\ \left\langle R^1 \right\rangle_{\hat{p}^2} &= \sum_{x^2_{1:T}} \hat{p}^2(x^2_{1:T}|x^1_0, x^2_0) R(x^1_{1:T}, x^2_{1:T}) = \sum_{t=1}^{T} \sum_{x^2_t} \hat{p}^2(x^2_t|x^1_0, x^2_0) R^1(x^1_t, x^2_t) \end{aligned}$$

Two agents cooperative games

How do we choose the opponent model?

When the problem is symmetric:
- agents are identical (same states, same q)
- the reward is symmetric $R^1(x^1, x^2) = R^2(x^2, x^1)$

one can use a recursive argument leading to an infinite sequence of nested beliefs

Agent 1:
- assumes an initial opponent model $p_0^2(x_{1:T}^2|x_0^1, x_0^2)$
- computes its optimal behaviour $p^1(x_{1:T}^1|x_0^1, x_0^2)$
- reasons, that agent 2 could have done the same.
- assumes new opponent model $p_1^2(x_{1:T}^2|x_0^1, x_0^2) = p^1(x_{1:T}^2|x_0^2, x_0^1)$
- computes its optimal behaviour p^1 against p_1^2
- ...

Two agents cooperative games

$$
\begin{aligned}
C^1(p_{k+1}|x_0^1, x_0^2) &= KL(p_{k+1}||q) - \left\langle R^1 \right\rangle_{p_{k+1}, p_k} \\
p_{k+1}(x_{1:T}^1|x_0^1, x_0^2) &= \frac{1}{Z} q(x_{1:T}^1|x_0^1) \exp\left(\left\langle R^1 \right\rangle_{p_k}\right)
\end{aligned}
$$

The infinite recursion leads to a fixed point equation with solution $p_\infty(x_{1:T}^1|x_0^1, x_0^2) = \lim_{k\to\infty} p_{k+1}(x_{1:T}^1|x_0^1, x_0^2)$, where both agents play the same.

Stag hunt game

	Stag	Hare
Stag	4,4	1,3
Hare	3,1	3,3

Get a Hare for yourself or a Stag together.

Two Nash equilibria:
if opponent plays Stag, I play Stag
if opponent plays Hare, I play Hare

Model for human and animal cooperation:
- slime molds can stick together to reproduce
- orcas can catch large schools of fish

Static stag hunt game

$x = \pm 1$ denotes Stag or Hare. Reward matrix $R(x^1, x^2)$:

	1	-1
1	4,4	1,3
-1	3,1	3,3

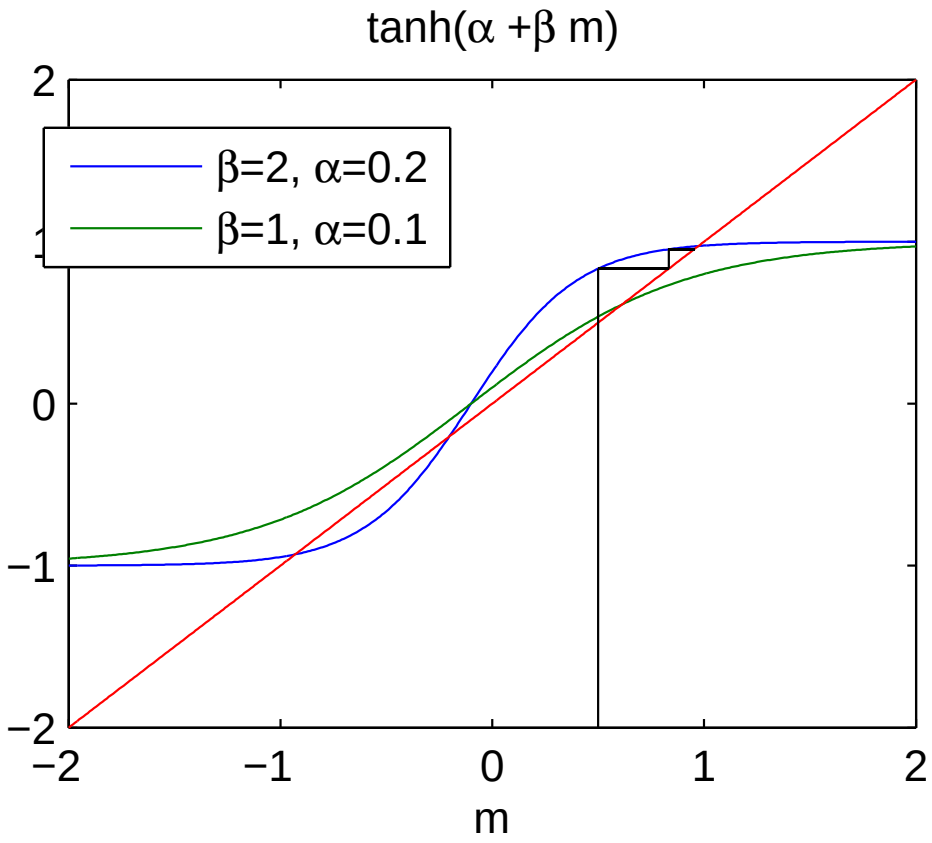
The game is only played once, ie. $T = 1$.

There is no dependence on the current state, so that $q(x_{1:T}|x_0) = 1$.

We can express $p_k(x)$ in terms of its expectation value m_k as $p_k(x) = \frac{1}{2}(1 + m_k x)$.

$$
\begin{aligned}
m_{k+1} &= \tanh\left(\frac{1}{2}\sum_{x'}(1 + m_k x')\left(R(1, x') - R(-1, x')\right)\right) = \tanh(\alpha + \beta m_k) \\
\alpha &= \frac{1}{2}(R(1,1) + R(1,-1) - R(-1,1) - R(-1,-1)) \\
\beta &= \frac{1}{2}(R(1,1) - R(1,-1) - R(-1,1) + R(-1,-1))
\end{aligned}
$$

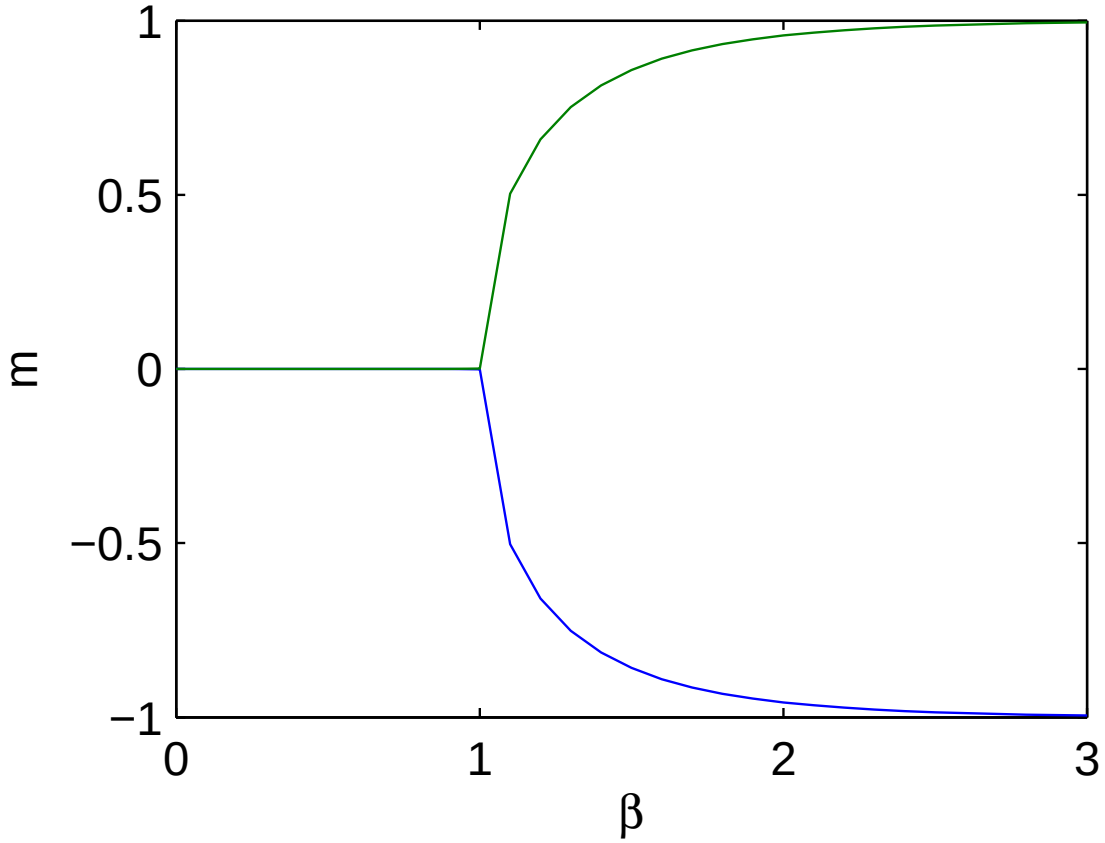
Static stag hunt game



$m_{k+1} = \tanh(\alpha + \beta m_k)$ versus m_k.

For small β there is a unique solution.
For large β there are two solutions, and dependence on initial conditions.

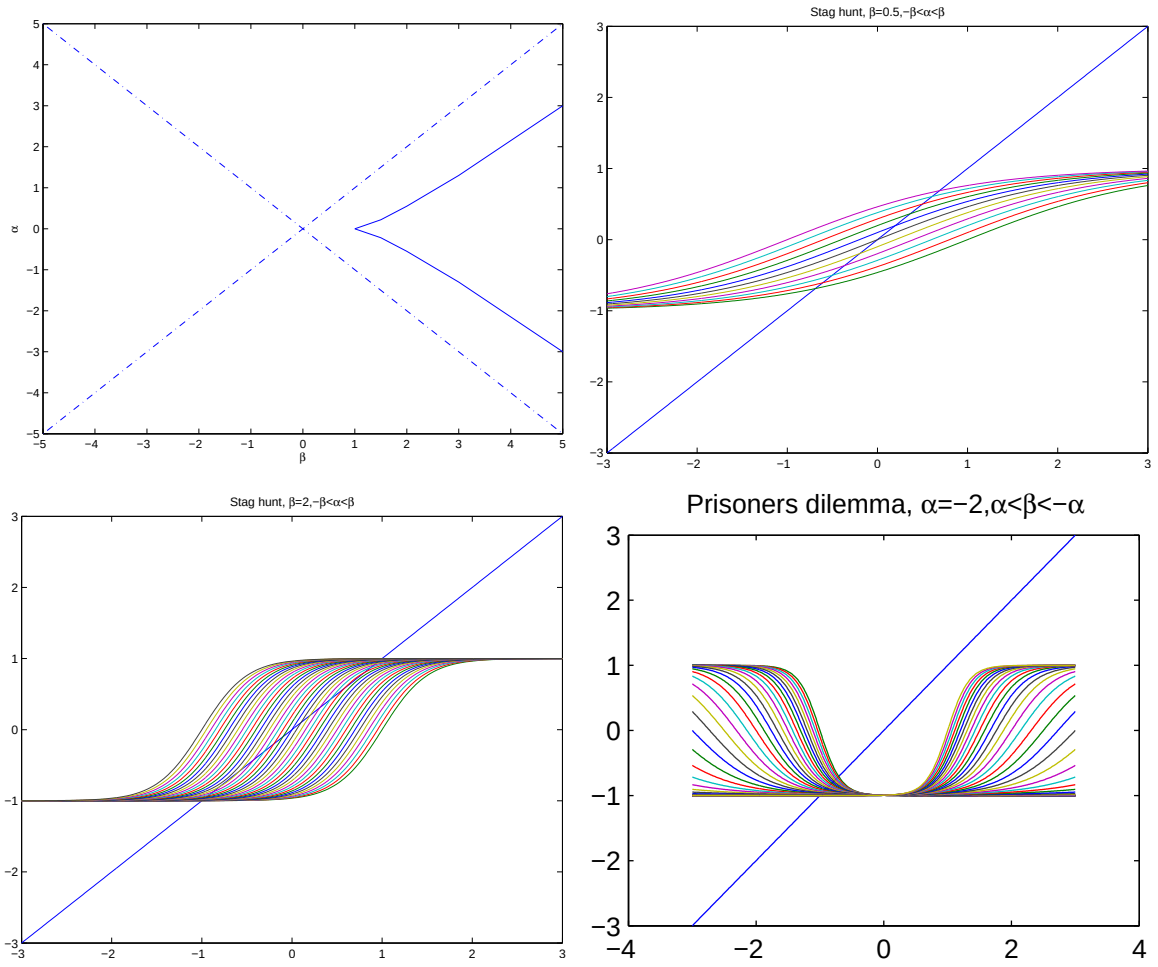
Static stag hunt game



m versus β for $\alpha = 0$.

Static stag hunt game

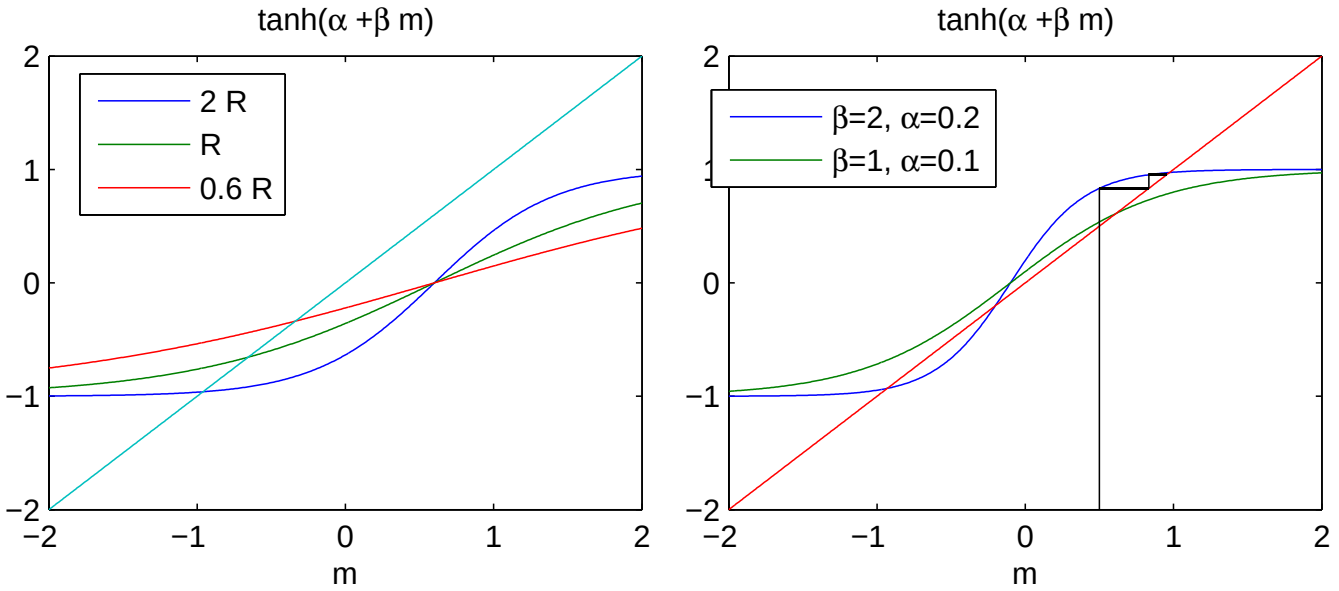
The two Nash equilibria imply $\beta > 0, -\beta < \alpha < \beta$.



Stag hunt game has local minima. Other games, such as Prisoners Dilemma, not.

Static stag hunt game

Thus:
- In many settings of the game there are two solutions $m \approx \pm 1$.
- The solution $m \approx 1$ is always better than the solution $m \approx -1$.
- The solution that is found depends on the initial assumption about the opponent
- Smaller rewards yield more likely cooperation



Dynamic stag hunt game

Optimal control is computed by backwards message passing:

$$
\begin{aligned}
C^1(p_{k+1}|x_0^1, x_0^2) &= KL(p_{k+1}||q) - \left\langle R^1 \right\rangle_{p_{k+1}, p_k} \\
p_{k+1}(x_{1:T}^1|x_0^1, x_0^2) &= \frac{1}{Z} q(x_{1:T}^1|x_0^1) \exp\left(\left\langle R^1 \right\rangle_{p_k}\right)
\end{aligned}
$$

$\left\langle R^1 \right\rangle_{p_k}$ is the expected future reward of agent 1's trajectory $x_{1:T}^1$ when agent 2 acts according to $p_k(x_{1:T}^2|x_0^1, x_0^2)$. It can be computed as a prediction:

$$
\begin{aligned}
\left\langle R^1 \right\rangle_{p_k}(x_{1:T}^1) &= \sum_{x_{1:T}^2} p_k(x_{1:T}^2|x_0^1, x_0^2) R(x_{1:T}^1, x_{1:T}^2) \\
&= \sum_{t=1}^{T} \sum_{x_t^2} p_k(x_t^2|x_0^1, x_0^2) R_t(x_t^1, x_t^2) = \sum_{t=1}^{T} \left\langle R_t^1 \right\rangle (x_t^1)
\end{aligned}
$$

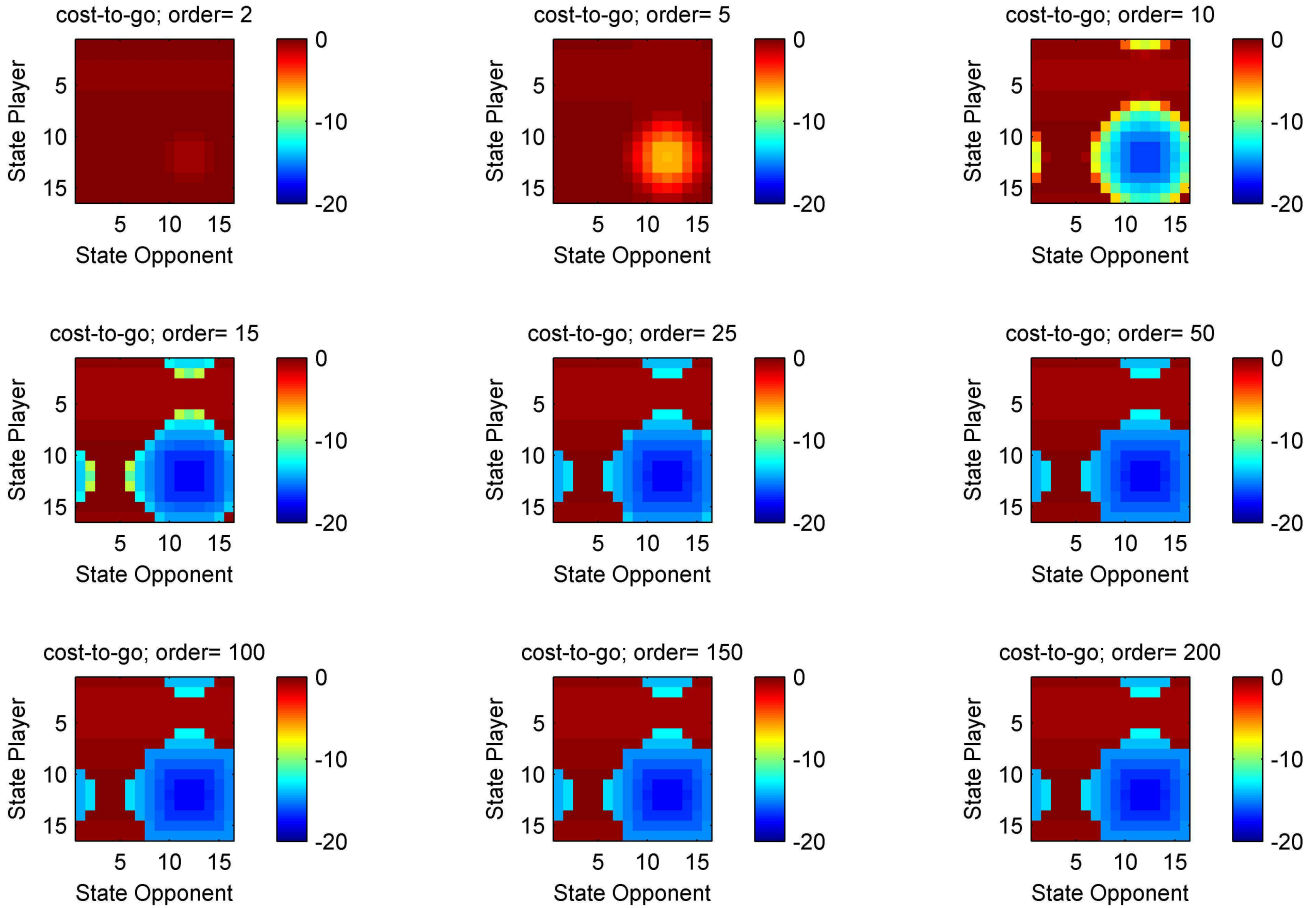
Dynamic stag hunt game

Initialize $p_0(x_{1:T}|x_0^1, x_0^2) = q(x_{1:T}|x_0^1, x_0^2)$ a random walk.

For $k = 0, 1, 2, \ldots$
- Predict $\left\langle R_t^1 \right\rangle_{p_k} (x_t^1), t = 1, \ldots, T$
- Compute $p_{k+1}(x_{1:T}^1|x_0^1, x_0^2)$

End

Dynamic stag hunt game



$T = 20, R_{\text{Stag}} = 0.1, R_{\text{Hare}} = 0.01, x_{\text{Stag}} = 12, x_{\text{Hare}} = 4.$ Brown=Hare; Blue=Stag

Discussion and future work

KL/Path integral control theory as a computationally feasible approach to
- agent coordination Firemen: end cost only; approx inference
- games with delayed rewards Stag hunt: ficticious play as a variational approximation

Factorization over agents as variational solution

$$KL(p|q) - \langle R \rangle \qquad p(x^1_{1:T}, x^2_{1:T}) = p_1(x^1_{1:T}) p_2(x^2_{1:T})$$

Future:

Repeated games:
- learning opponent behavior based on actual play
- POMDP or dual control setting; learning p or R?

Discussion and future work

Stochastic control theory predicts phase transition
- Optimal control is qualitatively different for high and low noise
- Delayed choice behavior
- Stat mech description and approximate inference

Robot application

See my webpage for papers www.snn.ru.nl/~bertk

Acknowledgements

Julian Tramper, Stan Gielen on delayed choice experiments

Stijn Tonk on agent opponent modeling.

Probabilistic methods for robotics and control (NIPS Dec 12, 2009)

	X	Y
X	45,45	0,35
Y	35,0	40,40

FIGURE 1.—Game $2R$.

	X	Y
X	45,45	0,40
Y	40,0	20,20

FIGURE 2.—Game R.

	X	Y
X	45,45	0,42
Y	42,0	12,12

FIGURE 3.—Game $0.6R$.

$$\alpha/\beta = -3/5,\ \beta = 25, 12.5, 7.5$$

8 cohorts, 8 subjects per cohort. Each cohort plays one of the three games, 75 times. Subjects are randomly paired within a cohort.

BSH 2001, Econometrica

CONTINGENCY TABLE I

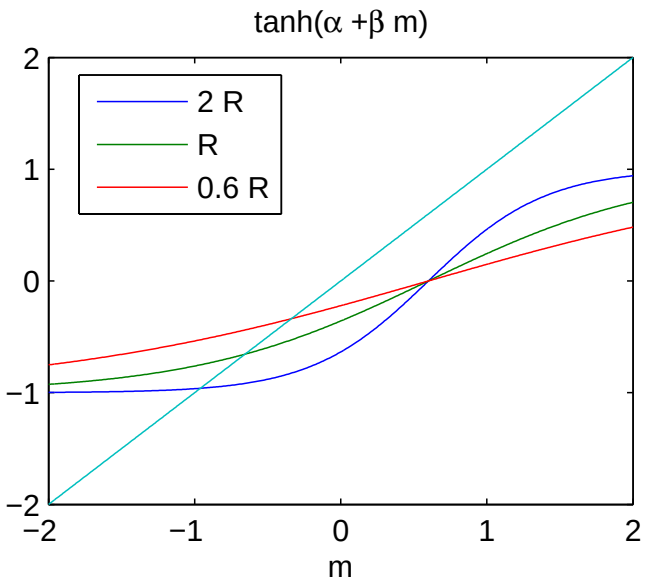
TREATMENT BY PERIOD 1 SUBJECT CHOICE

	X	Y	Total
$0.6R$	41 (0.64)	23 (0.36)	64 (1.00)
R	45 (0.70)	19 (0.30)	64 (1.00)
$2R$	34 (0.53)	30 (0.47)	64 (1.00)
Total	120 (0.63)	72 (0.37)	192 (1.00)

CONTINGENCY TABLE II

TREATMENT BY PERIOD 75 SUBJECT CHOICE

	X	Y	Total
$0.6R$	28 (0.44)	36 (0.56)	64 (1.00)
R	16 (0.25)	48 (0.75)	64 (1.00)
$2R$	3 (0.05)	61 (0.95)	64 (1.00)
Total	47 (0.24)	145 (0.76)	192 (1.00)



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.

Reversing time

Let $\rho(y,\tau|x,t)$ describe a diffusion process defined by the Fokker-Planck equation

$$\partial_\tau \rho = H^\dagger \rho \tag{1}$$

with $\rho(y,t|x,t) = \delta(y-x)$.

Define

$$A(x,t) = \int dy \rho(y,\tau|x,t)\Psi(y,\tau).$$

It is easy to see by using the equations of motions for Ψ and ρ 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 $\tau = t_f$ yields $A(x,t) = \int dy \rho(y,t_f|x,t)\Psi(y,t_f)$. Thus,

$$\Psi(x,t) \quad = \quad \int dy \rho(y,t_f|x,t) \exp(-\phi(y)/\nu) \tag{2}$$

MC sampling

The diffusion equation

$$\partial_\tau \rho = -\frac{V}{\nu}\rho - \partial_y(f\rho) + \frac{1}{2}\nu\partial_y^2\rho \tag{3}$$

can be sampled as

$$\begin{aligned} dx &= f(x,t)dt + d\xi \\ x &= x + dx, \quad \text{with probability } 1 - V(x,t)dt/\nu \\ x_i &= \dagger, \quad \text{with probability } V(x,t)dt/\nu \end{aligned} \tag{4}$$

MC sampling

We can estimate

$$\Psi(x,t) \quad = \quad \int dy \rho(y,t_f|x,t) \exp(-\phi(y)/\nu) \tag{5}$$

by computing N trajectories $x_i(t \rightarrow t_f), i = 1, \ldots, N$.

Then, $\Psi(x,t)$ is estimated by

$$\hat{\Psi}(x,t) = \frac{1}{N} \sum_{i \in \text{alive}} \exp(-\phi(x_i(t_f))/\nu) \tag{6}$$

where 'alive' denotes the subset of trajectories that do not get killed along the way by the † operation.

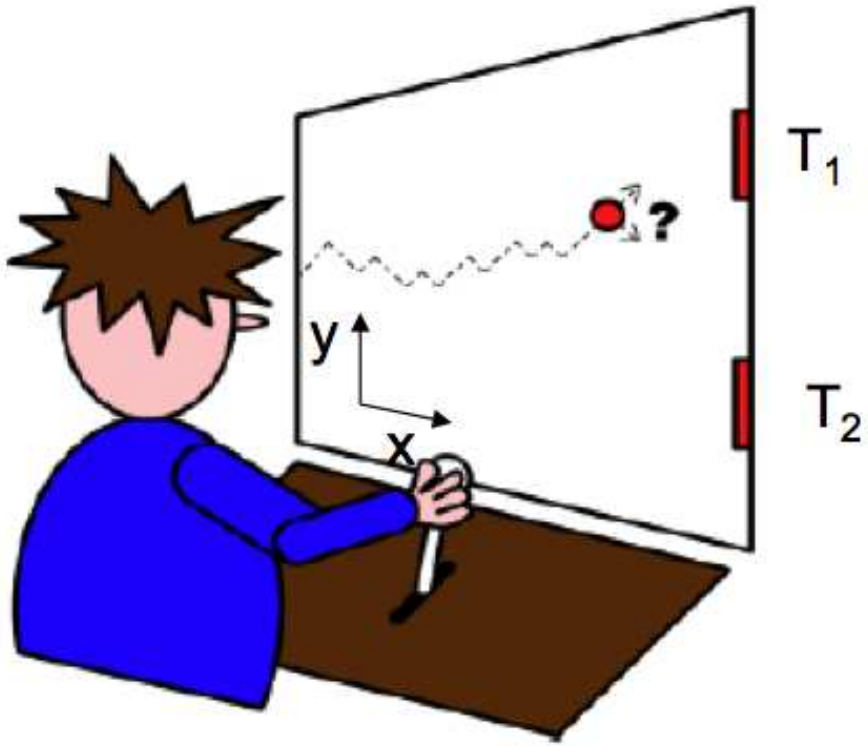
Experiment

Cursor position (x(t),y(t))

$$x(t) = vt$$

$$\frac{dy}{dt} = u(t) + \xi(t)$$

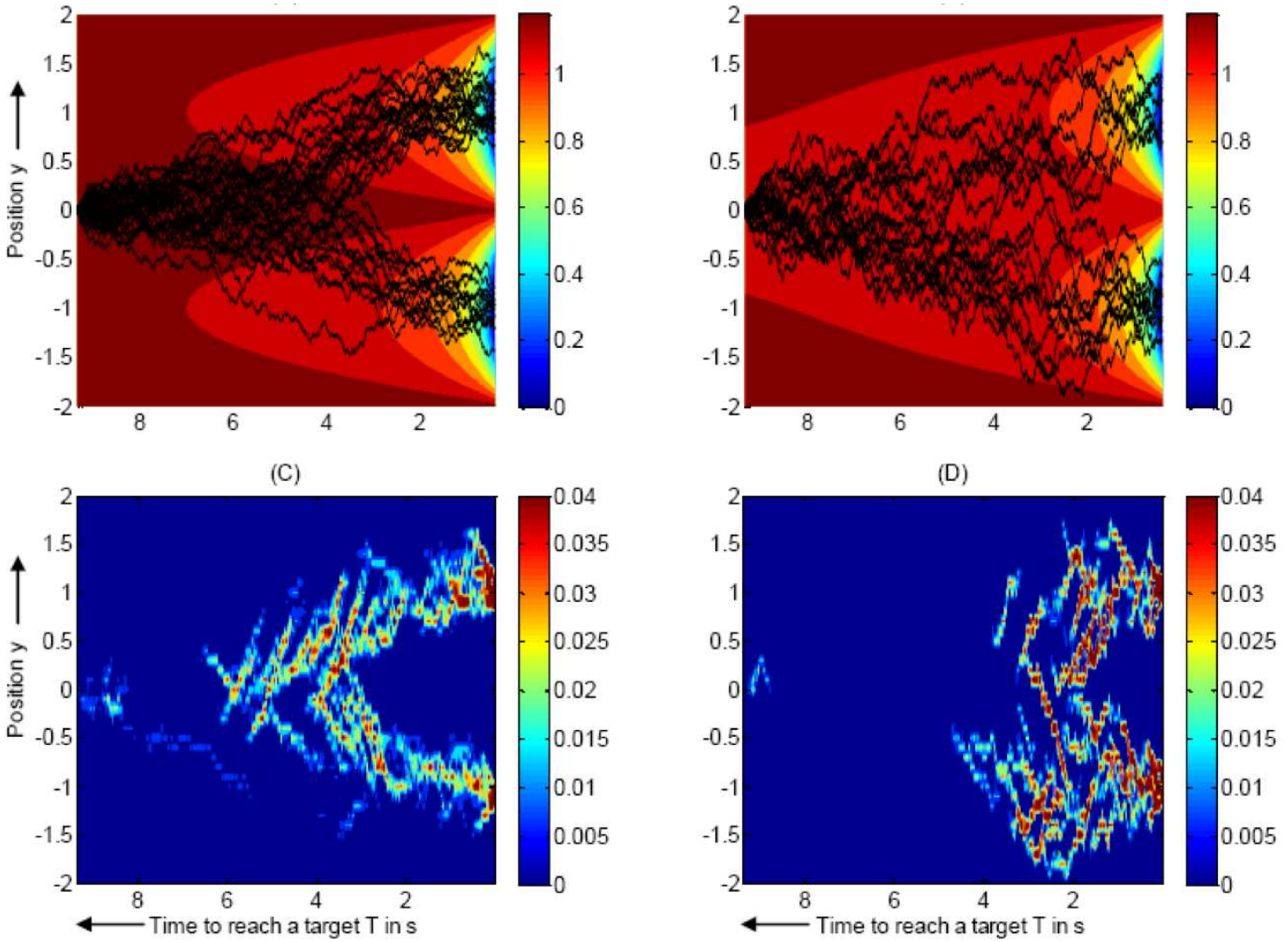
with $\left\langle \xi^2(t) \right\rangle = \sigma^2$



Instruction to the subject:

make sure that the cursor hits one of the two targets T_1 or T_2

Results



Low noise

High noise

Conclusions

Stochastic optimal control theory predicts that
- Delayed choice is an optimal strategy
- This behavior is also observed in humans
- Results are preliminary and need further quantification