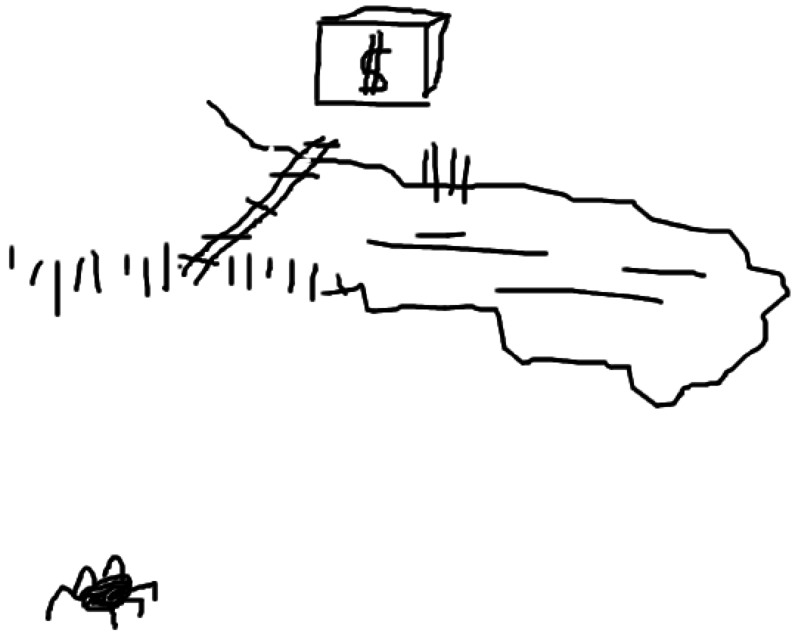
KL control theory and decision making under uncertainty

Bert Kappen
SNN Radboud University
Nijmegen

with Stijn Tonk

December 12, 2009

Stochastic optimal control theory



optimal solution is noise dependent

computation is intractable

KL control theory

Linear control theory (K 2005)

- continuous state and time, Gaussian noise, arbitrary reward and dynamics, additive control
- log transform linearizes Bellman equation (Schrödinger equation, Fleming)
- optimal cost-to-go as a free energy

$$J(x) = -\nu \log \sum_{x_{dt:T}} \exp\left(-S(x_{dt:T})/\nu\right)$$

- phase transitions
- graphical model (approximate) inference

Discrete state & time case using KL (Todorov 2006)

Relation between the two approaches (K et al. arxiv)

Opponent modeling

Agents successfull behavior depends on adequate model of environment and other agents behavior.

- dialogue maintenance
- man-machine interfaces
- team play

Either cooperative or antagonistic

Today's talk

Approximate inference

KL control theory

Opponent modeling, nested beliefs or levels of sophistication

- KL control for agents; opponent models
- stag hunt game

Conclusions

Approximate inference

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

$$
\begin{aligned}
KL(p||\exp(-E)) &= \sum_x p(x)\log\frac{p(x)}{\exp(-E(x))} \\
p^*(x) &= \mathrm{argmin}_p KL(p||\exp(-E)) \\
KL(P^*||\exp(-E) &= -\log Z
\end{aligned}
$$

Approximate inference:
- approximate KL
- restrict minimization to tractable class of p

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

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

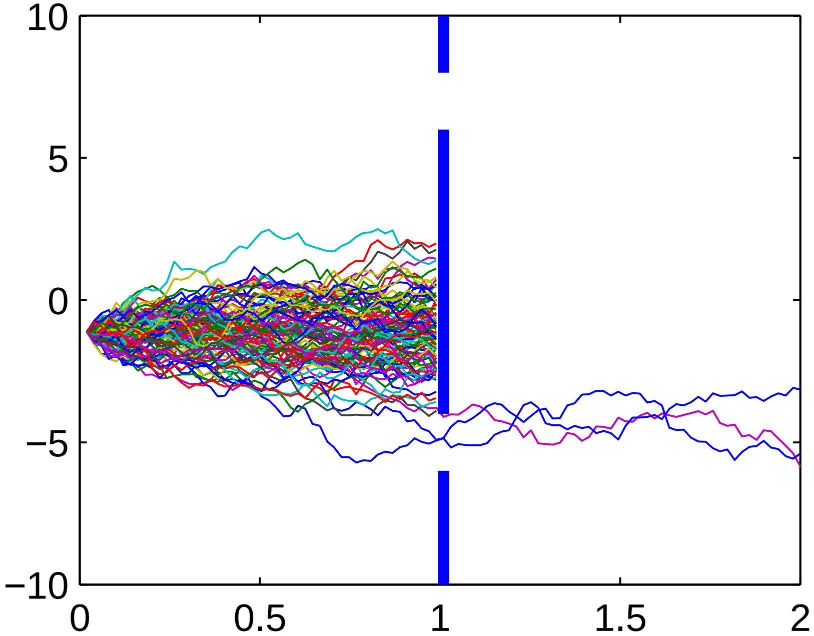
$q(x_{t+1}|x_t)$ implements the allowed moves.

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

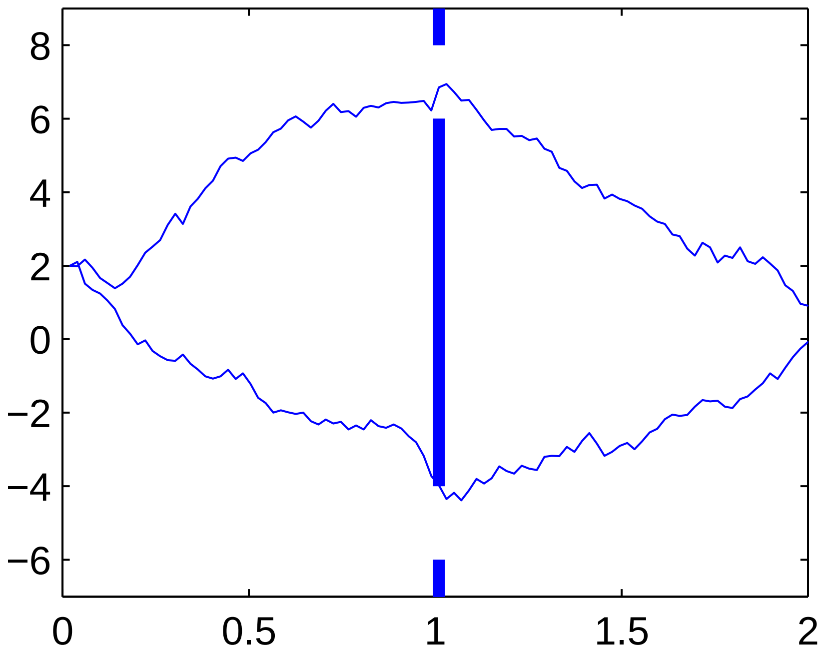
The KL control problem is to find the probability distribution $p(x_{1:T}|x_0)$ that minimizes

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

KL control theory



(a) Sample paths under q



(b) Sample paths under p

KL control theory

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

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

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

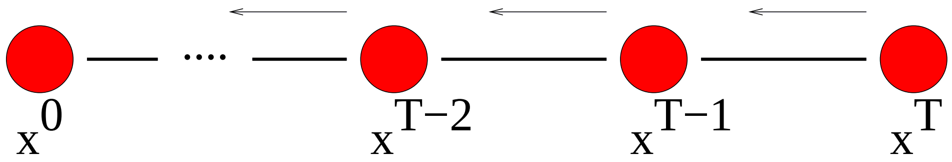
NB: $Z(x_0)$ is an integral over paths.

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.



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

KL control theory

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

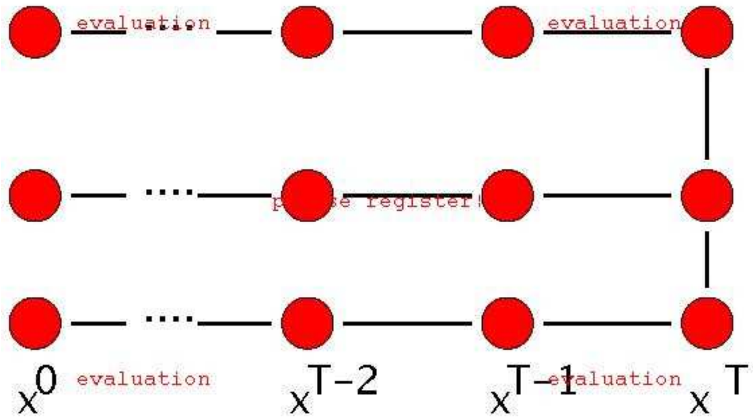
Dynamics: $p^t_{xy}(\pi)$ Cost: $C(\pi^{0:T}) = -\langle R \rangle$	$\rightarrow$ DP $\rightarrow$	Bellman Equation
$\downarrow$		$\downarrow$
restricted class		approximate J
$\downarrow$		$\downarrow$
Free dynamics: q^t_{xy} $C = KL(p\|\|q\exp(R))$	$\rightarrow$ approx inference $\rightarrow$	Optimal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

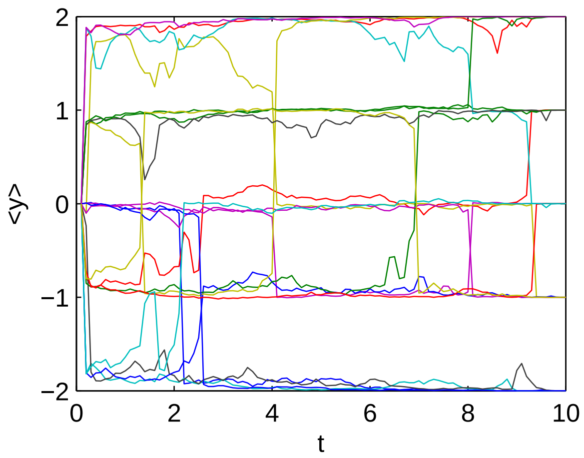
Optimal solution:

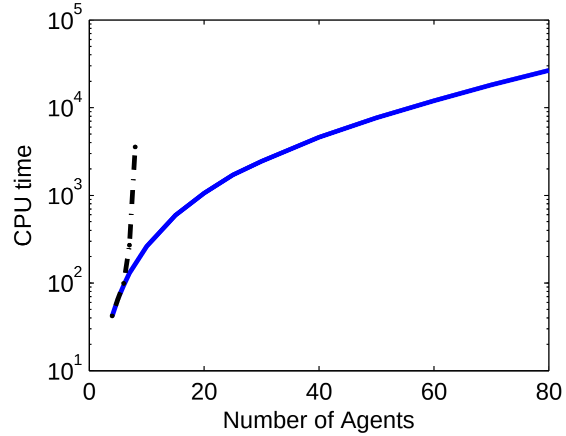
$$p(x^{1:T}|x^0) \quad = \quad \frac{1}{Z} q(x^{1:T}|x^0) \exp(R(x^{0:T}))$$

Intractable, but standard approximate inference problem.

Approximate inference for agent coordination using BP

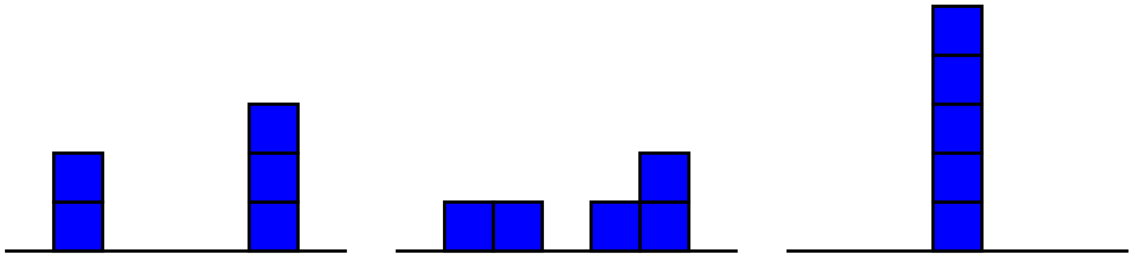


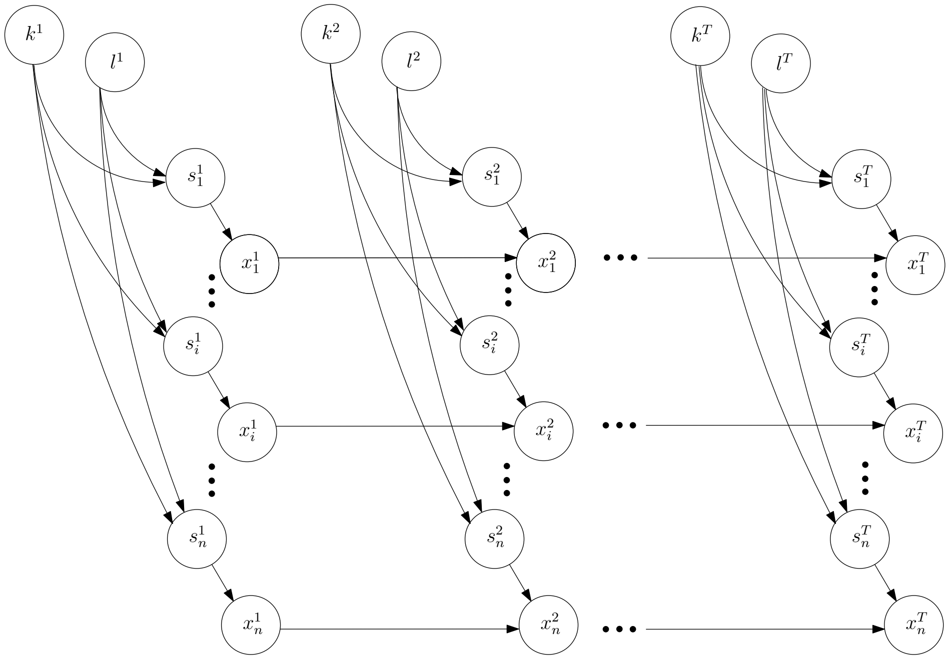


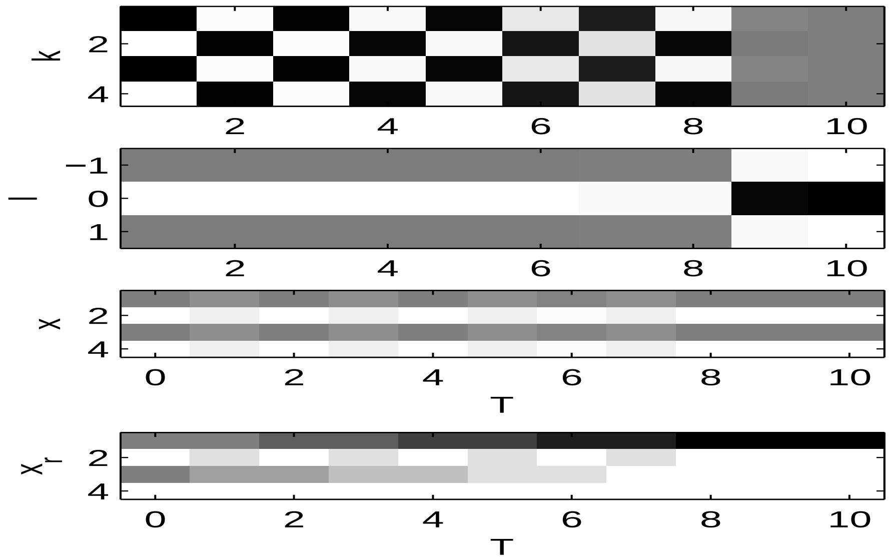


Broek et al. 2006

Approximate inference for stacking blocks using CVM







Double loop inside!

Kappen et al. arxiv.org

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.

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

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

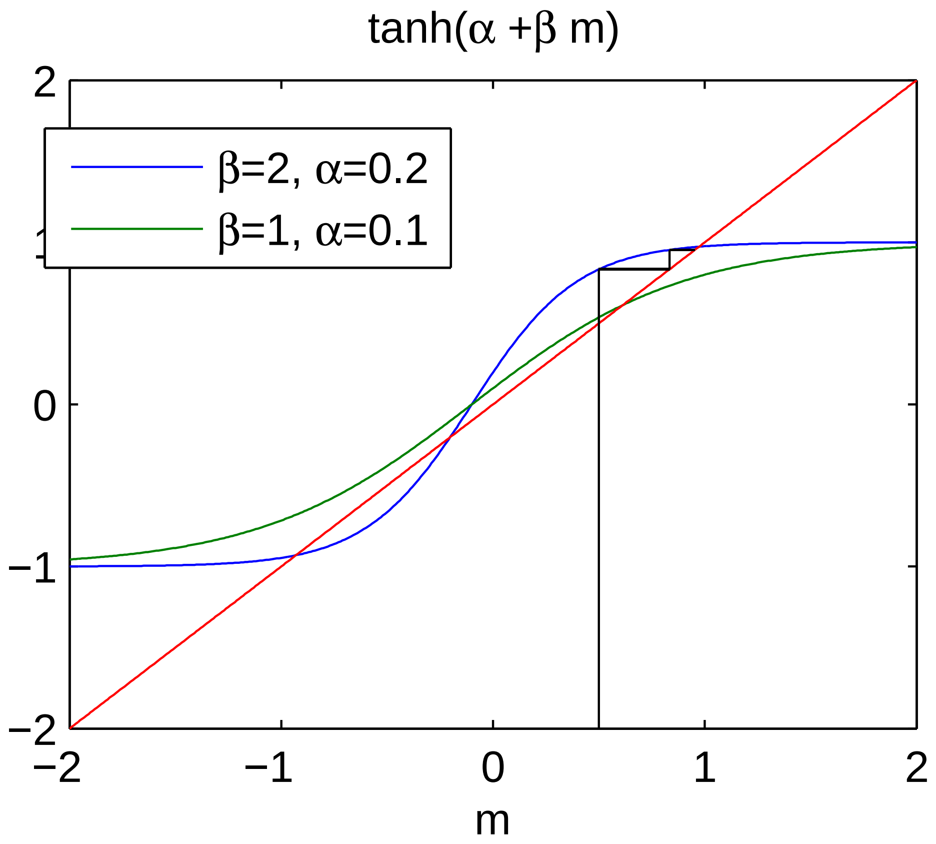
The game is only played once, ie. $T = 1$.

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 $p_k(x) = \frac{1}{2}(1 + m_k x)$.

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

Static stag hunt game

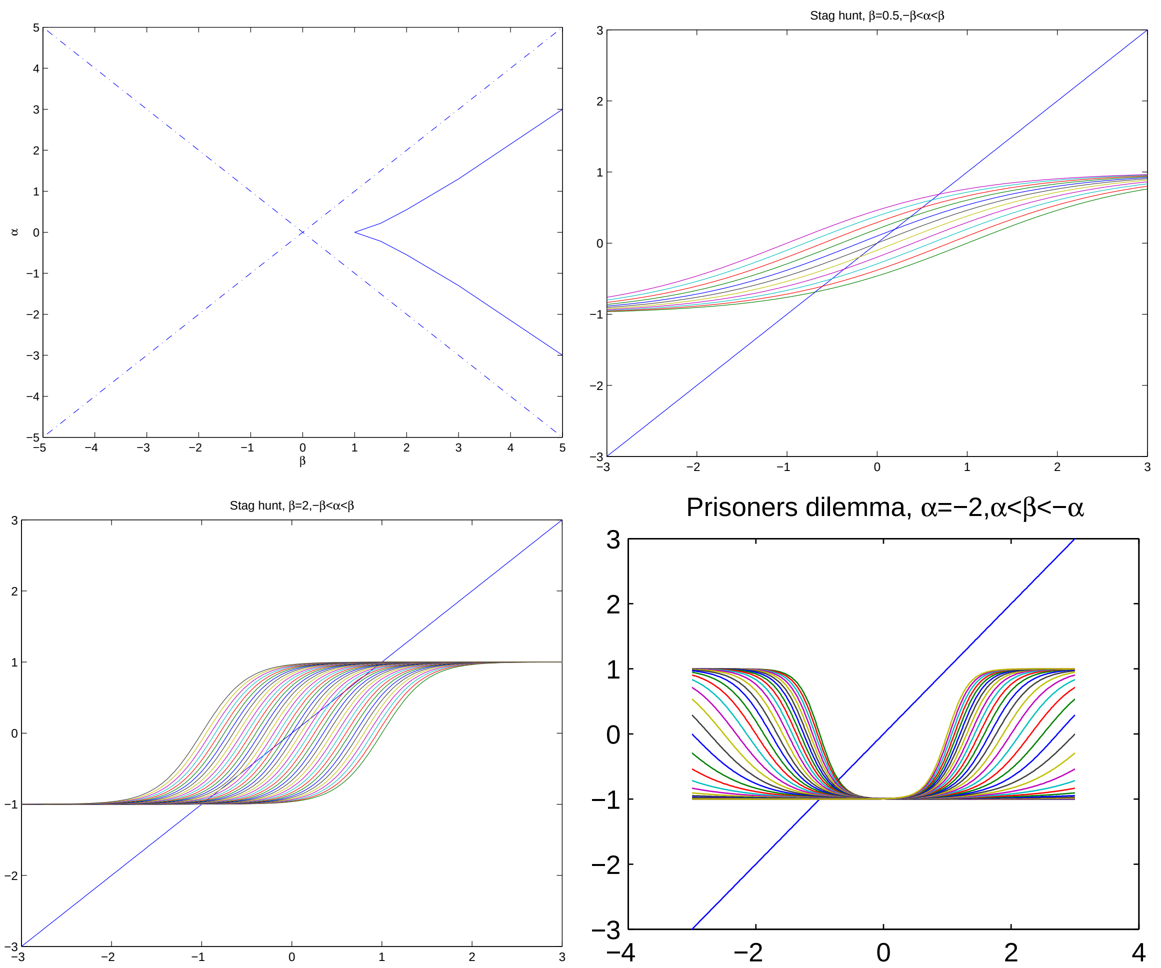


$$m_{k+1} = \tanh(\alpha + \beta m_k) \text{ versus } m_k.$$

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

Static stag hunt game

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



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

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:

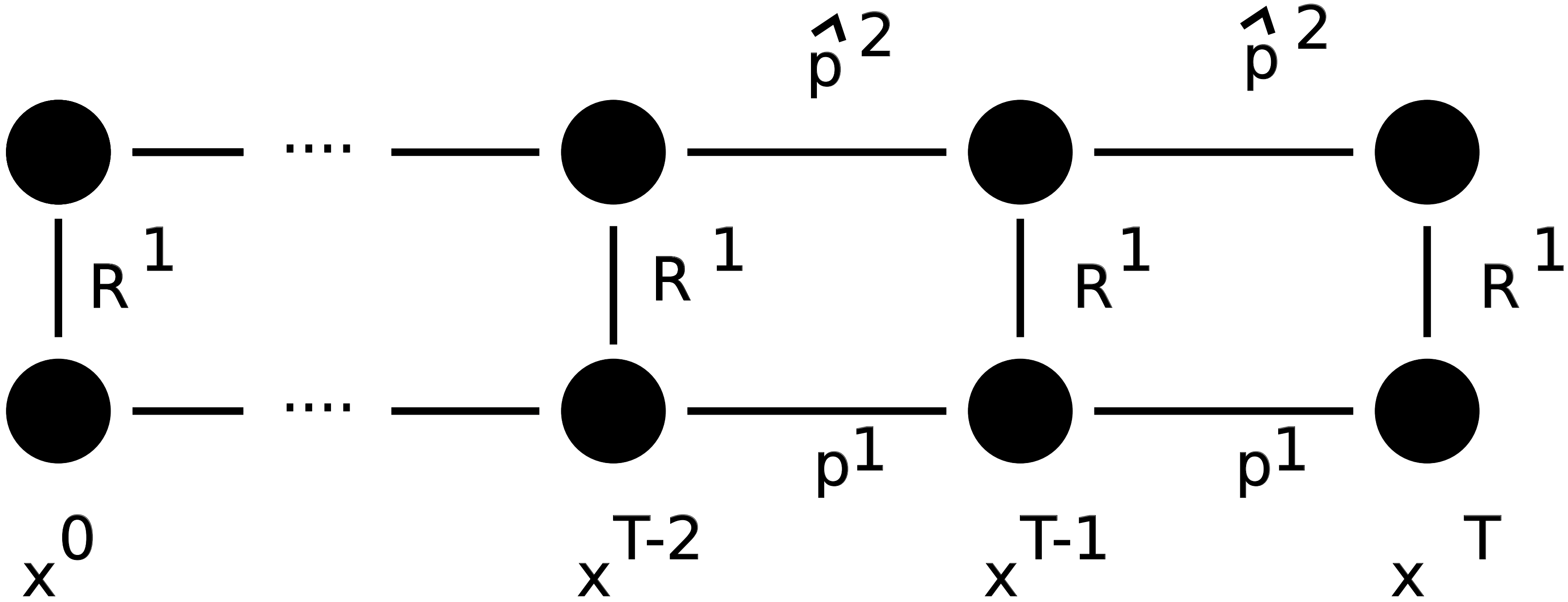
$$
\begin{aligned}
\left\langle R^1 \right\rangle_{p_k}(x_{1:T}^1) &= \sum_{x_{1:T}^2} p_k(x_{1:T}^2|x_0^1, x_0^2) R(x_{1:T}^1, x_{1:T}^2) \\
&= \sum_{t=1}^{T} \sum_{x_t^2} p_k(x_t^2|x_0^1, x_0^2) R_t(x_t^1, x_t^2) = \sum_{t=1}^{T} \left\langle R_t^1 \right\rangle (x_t^1)
\end{aligned}
$$

Dynamic stag hunt game

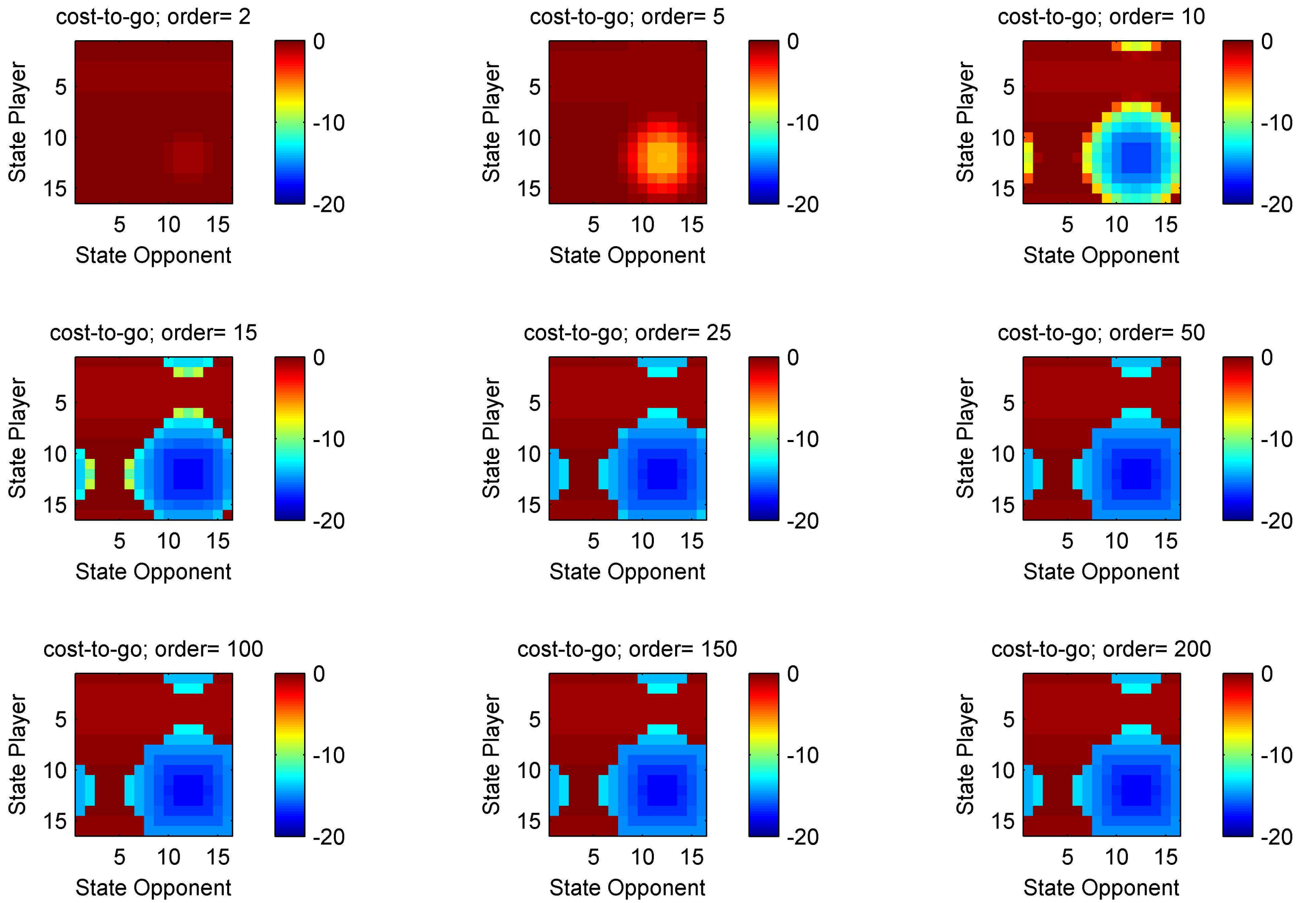
Initialize $p_0(x_{1:T}|x_0^1, x_0^2) = q(x_{1:T}|x_0^1, x_0^2)$ a random walk.

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

Conclusions

Path integrals for non-LQG control problems

- relating inference and control
- connection to other work presented here

Efficient approximations through

- particle filters, MCMC
- deterministic approximations

Main research issues:

- partial observability
- (reinforcement) learning

Conclusions

Nested beliefs recursion ('sophistication')

- example of non-trivial multi-agent reasoning
- extension to moving targets (poster)

Main research issues:

- antagonist or non-symmetric case
- learning based on actual play (POMDP setting)