

Approximate inference methods for stochastic optimal control theory.

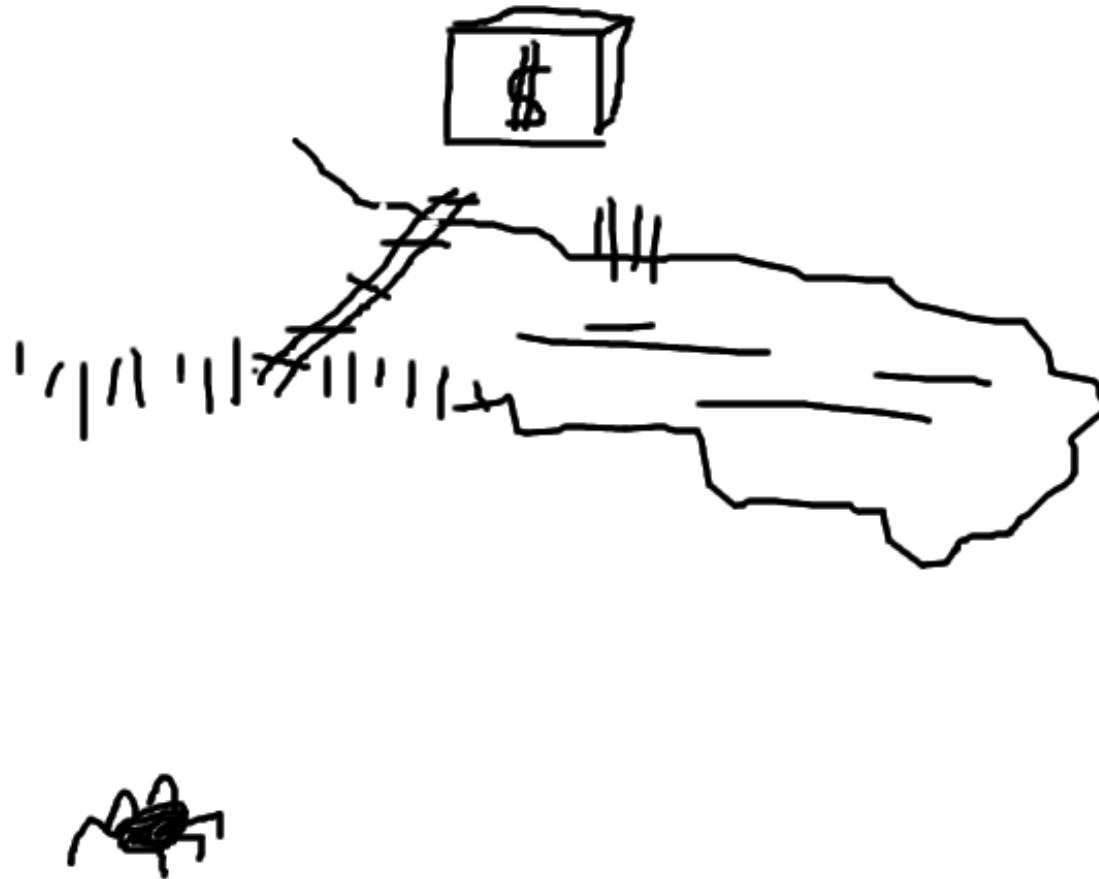
Bert Kappen, Vicenc Gomez
SNN Radboud University
Nijmegen

Manfred Opper
TU Berlin

December 13, 2008



Stochastic optimal control theory



Optimal solution is noise dependent and intractable

Stochastic optimal control theory

$x = 1, \dots, N$ denote states of the system, x^t is the state at time t .

$p_{xy}^t(u)$ is a Markov transition probability from x to y at time t under control u .

$p(x^{1:T} | x^0, u^{0:T-1})$ is the probability to observe the trajectory $x^{1:T}$ given initial state x^0 and control trajectory $u^{0:T-1}$.

If the system in state x takes action u there is an associated cost $R(x, u)$. The control problem is to find the sequence $u^{0:T-1}$ that minimizes:

$$C(x^0, u^{0:T-1}) = \sum_{x^{1:T}} p(x^{1:T} | x^0, u^{0:T-1}) \sum_{t=0}^{T-1} R(x^t, u^t) = \left\langle \sum_{t=0}^{T-1} R(x^t, u^t) \right\rangle$$

Stochastic optimal control theory

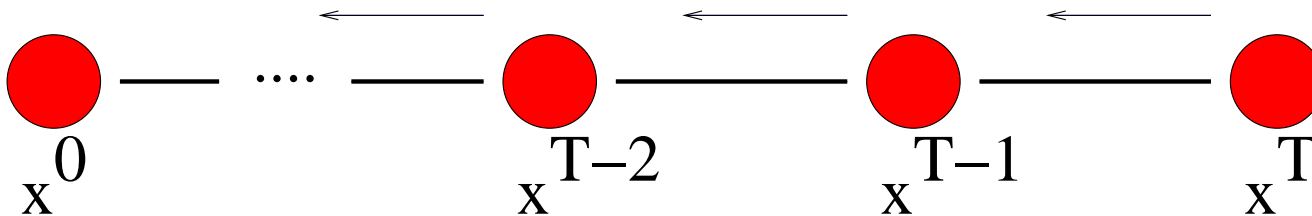
The optimal control is computed using the Bellman equation, which results from a dynamic programming argument.

For any intermediate t

$$J(T, x) = 0$$

$$\begin{aligned} J(t, x^t) &= \min_{u^{t:T-1}} \left\langle \sum_{s=t}^{T-1} R(x^s, u^s) \right\rangle \\ &= \min_{u^t} \left(R(x^t, u^t) + \sum_{x^{t+1}} p(x^{t+1} | x^t, u^t) J(t+1, x^{t+1}) \right) \end{aligned}$$

This is called the *Bellman Equation*, J is aka the value function.



Overview

Dynamics: $p_{xy}^t(u)$
Cost: $C(u^{0:T}) = \langle R \rangle$

→ DP →

Bellman Equation

↓
approximate J
↓
Optimal u

Overview

Dynamics: $p_{xy}^t(u)$
Cost: $C(u^{0:T}) = \langle R \rangle$

↓

restricted class

↓

Free dynamics: q_{xy}^t
 $C = KL(p||q \exp(-R))$

→ DP →

Bellman Equation

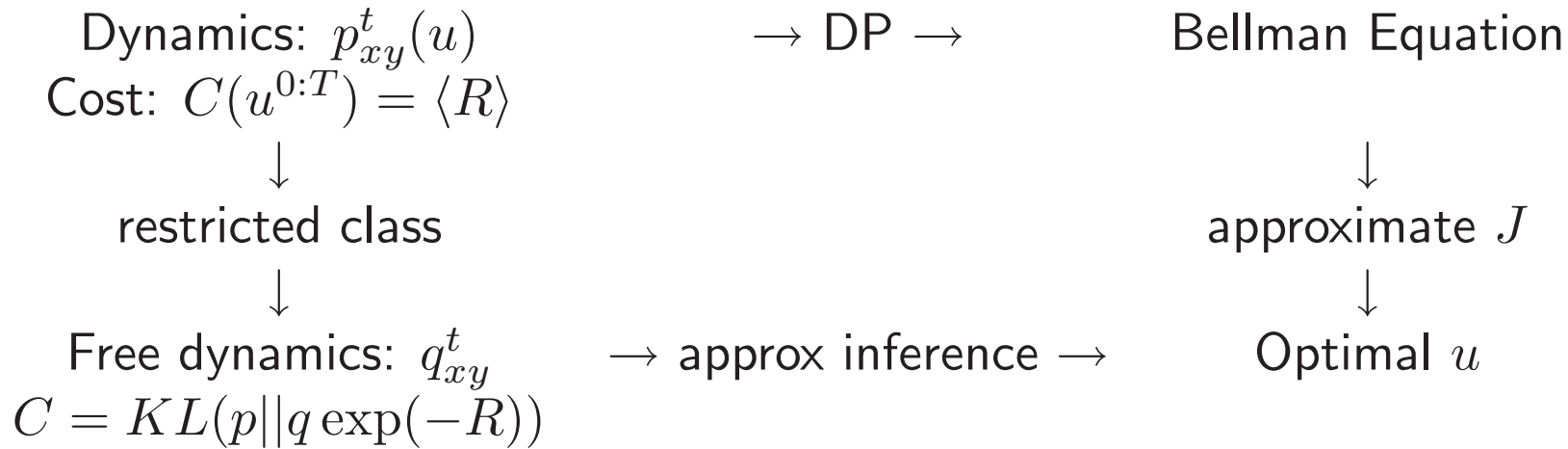
↓

approximate J

↓

Optimal u

Overview



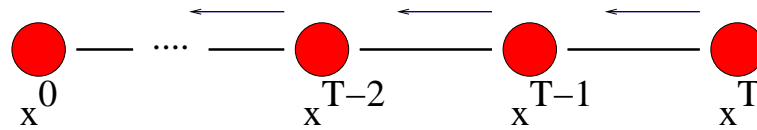
Optimal solution:

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

intractable, but standard approximate inference problem.

Comments

The proposed class of control problems was previously considered by Todorov, who identified for this class of problems the Bellman updates with the β messages for exact inference in a chain



This approach does not address the intractability issue.

The contribution of this paper is

- to write the control cost as a KL divergence

- to use approximate inference to compute the optimal control

The equivalence of certain types of control problems and inference problems is well-known for the linear quadratic Gaussian case (Kalman) and also for certain classes of non-linear problems (Kappen).

Outline of the talk

<L control

Relation to continuous path integral approach

A numerical example.

JT

double loop CVM

Summary and discussion

KL control theory

Instead of specifying p in terms of u and introducing a cost for u we minimize p directly.

optimal $u \leftrightarrow$ optimal p .

We assume the existence of a 'free' (uncontrolled) dynamics q and seek a controlled dynamics p such that

$$C(x^0, p) = KL(p||q) + \langle R \rangle = \sum_{x^{1:T}} p(x^{1:T}|x^0) \log \frac{p(x^{1:T}|x^0)}{r(x^{1:T}|x^0)}$$

$$r(x^{1:T}|x^0) = q(x^{1:T}|x^0) \exp \left(- \sum_{t=0}^T R(x^t, t) \right)$$

is minimized.

KL control theory

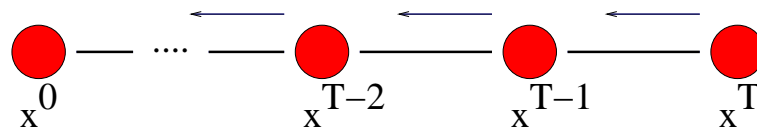
Minimizing the KL yields

$$p(x^{1:T}|x^0) = \frac{1}{Z(x_0)} r(x^{1:T}|x^0) \quad C(x^0, p) = -\log Z(x_0)$$

The optimal control at time t is given by

$$p(x^{t+1}|x^t) = \sum_{x^{t+2:T}} p(x^{t+1:T}|x^t) \propto q^t(x^{t+1}|x^t) \beta^{t+1}(x^{t+1})$$

with $\beta^t(x)$ the backward messages.



NB: $q(x^t|x^{t-1}) = q(x^t)$ yields $p(x^t|x^{t-1}) \propto q(x^t) \exp(-R(x^t))$.

Continuous space formulation

Consider

$$x^{t+1} = x^t + f(x^t, t) + u^t + \xi$$

The state x denotes an n -dimensional real vector. ξ is n -dimensional Gaussian noise with covariance matrix ν . u an n -dimensional control vector.

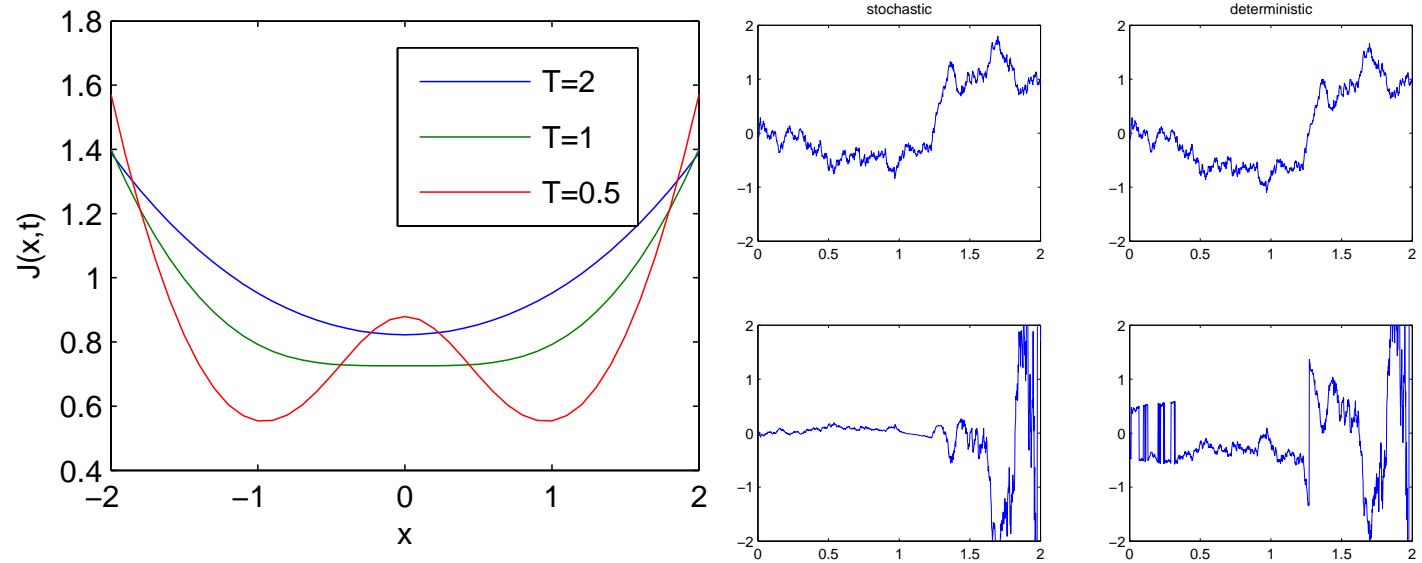
Previous results show that for a certain control cost, the optimal control computation is an inference problem:

The cost-to-go becomes a log path integral

This is special case of this KL control formulation and generalizes to discrete time.

$$\begin{aligned} p_{xy}^t(u) &= \mathcal{N}(y|x + f(x, t) + u(x, t), \nu) \\ q_{xy}^t &= \mathcal{N}(y|x + f(x, t), \nu) \\ KL(p||q) &= \sum_{x^{1:T}} p(x^{1:T}|x^0) \frac{1}{2} u(x^t, t)^T \nu^{-1} u(x^t, t) \end{aligned}$$

Continuous space formulation



Darts with EP

Graphical model inference

When x is high dimensional the control computation

$$p(x^{t+1}|x^t)$$

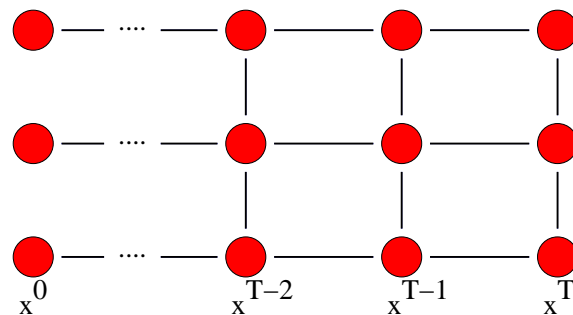
is intractable.

When one assumes that

the problem has some structure: $x = (x_1, \dots, x_n)$.

q factorizes: $q_{xy} = \prod_{i=1}^n q_i(y_i|x_i)$.

R has a sparse structure: $R(x) = \sum_{ij} R_{ij}(x_i, x_j)$.



We can approximate the computation of the marginal using standard methods.

Comments

Note, that in general

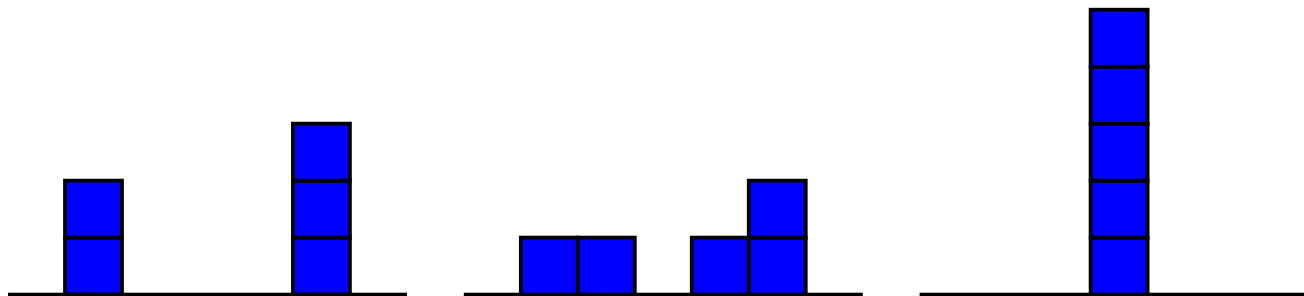
$$p(x_1^{t+1}, \dots, x_n^{t+1} | x_1^t, \dots, x_n^t) \neq \prod_{i=1}^n p(x_i^{t+1} | x_1^t, \dots, x_n^t).$$

This is the well-known coordination problem: the choice at one component affects the optimal choice at other components.

Exploiting graphical structure to obtain approximate solutions has been considered before in the RL community.

It has been observed that the quality of the solution is tightly linked to the sparsity structure of the problem, This is in agreement with the present approximate inference view.

A blocks world



m blocks and n possible block locations. $x_i^t = 0, \dots, m$ denotes the height of stack at location i at time t .

At iteration t , we move a block from location k^t to location $k^t + l^t$, with $l^t = -1, 0, 1$.

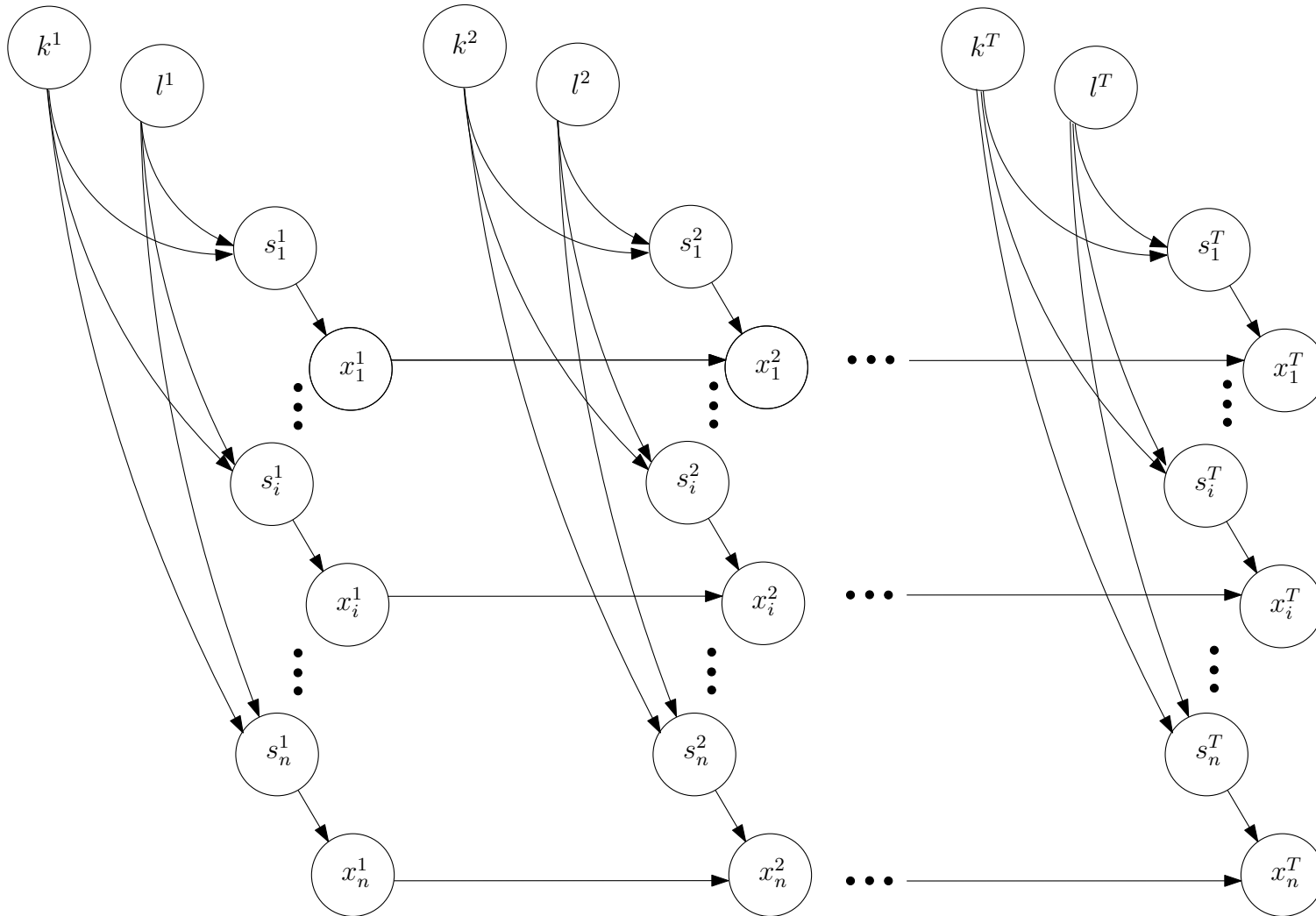
The Markov transition from x^t to x^{t+1} is a mixture over the values of k^t, l^t

$$q(k^t) = \mathcal{U}(1, \dots, n)$$

$$q(l^t) = \mathcal{U}(-1, 0, +1)$$

$$q(x^{t+1} | x^t) = \sum_{k^t, l^t} q(x^{t+1} | x^t, k^t, l^t) q(k^t) q(l^t)$$

A blocks world



A blocks world

Immediate cost is the entropy of the block configuration

$$R(x) = -\lambda \sum_i \frac{x_i}{m} \log \frac{x_i}{m}$$

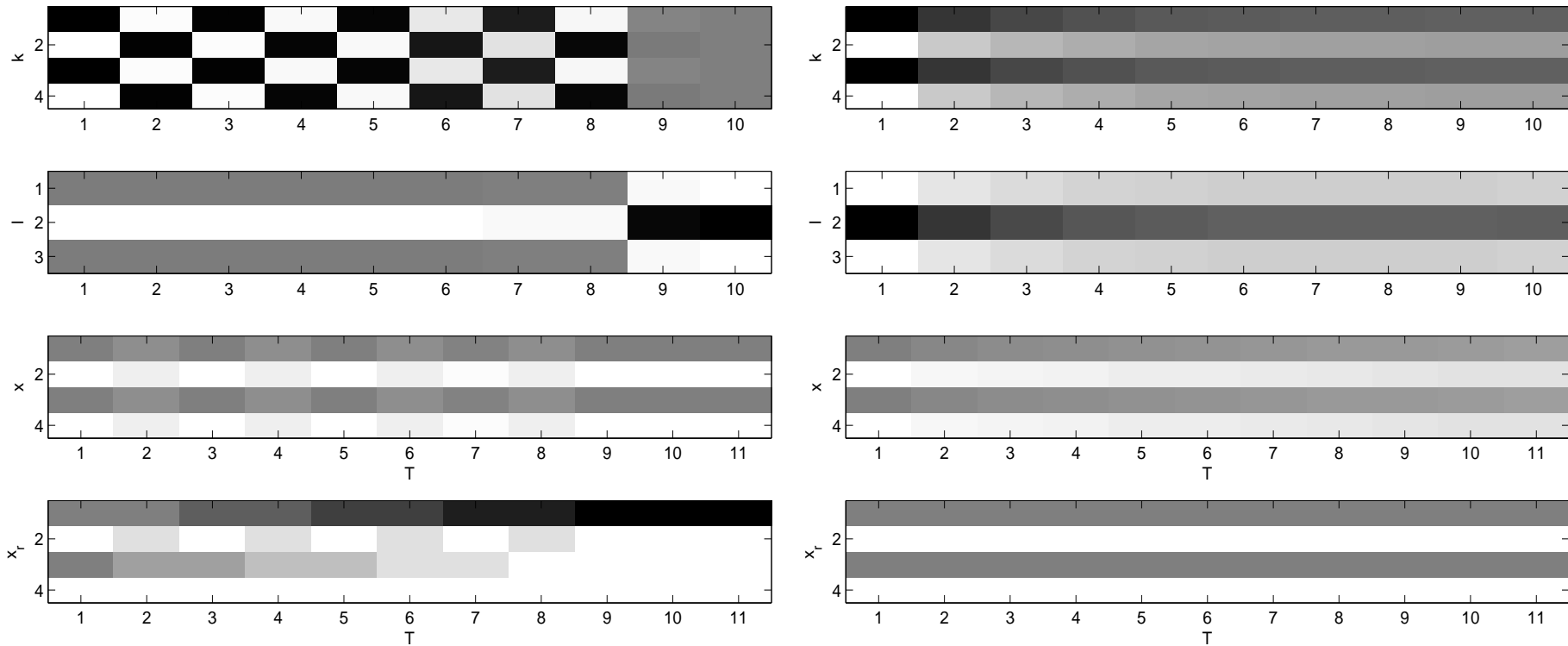
The minimum entropy solution puts all blocks on one stack.

The control problem is to minimize R over a future horizon T with a probability p that is as close as possible to q .

$$C = KL(p||q) + \langle R \rangle$$

λ small \leftrightarrow control is expensive \leftrightarrow p close to q .

A blocks world



$n = 4, m = 8, T = 10$ and $\lambda = 10$ (left) and $\lambda = 2$ (right) using exact inference. The blocks are initialized in two stacks of height 4. Each subfigure shows the marginals $p(k^t)$ (top), $p(l^t)$ (second) and $\langle x_i^t \rangle, i = 1, \dots, n$ (third) and the MAP solution (bottom) for $t = 1, \dots, T$ using a grey scale coding with white coding for zero and darker colors coding for higher values.

Cluster variation method

The cluster variation method replaces the probability distribution $p(x)$ by a large number of (possibly overlapping) probability distributions, each describing the interaction between a small number of variables.

$$p(x) \approx \{p_\alpha(x_\alpha), \alpha = 1, \dots\}$$

We approximate the control cost

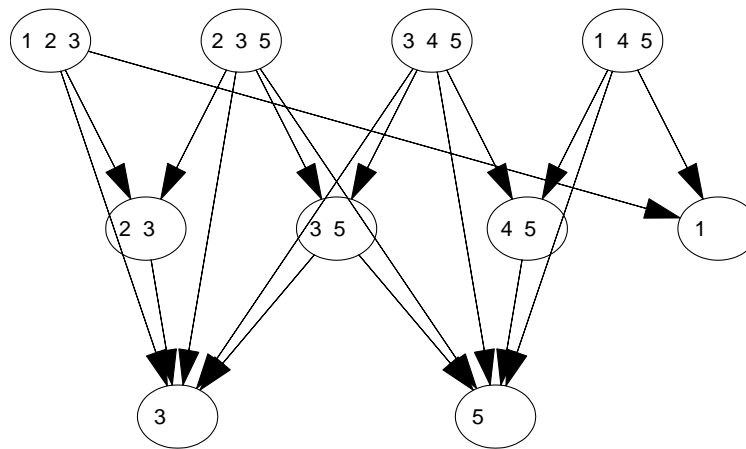
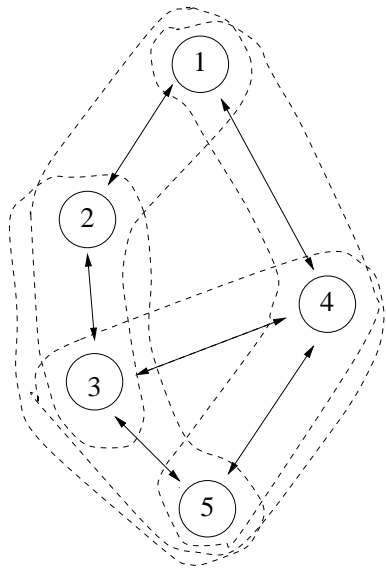
$$F(p) = \sum_x p(x) \log \frac{p(x)}{r(x)} \approx F_{\text{cvm}}(\{p_\alpha\})$$

and compute its minimum, subject to normalization and consistency constraints.

CVM Free Energy

$$F_{\text{cvm}} = \sum_{\alpha \in \mathcal{B}} \sum_{x_\alpha} p_\alpha(x_\alpha) \log \frac{p_\alpha(x_\alpha)}{\psi_\alpha(x_\alpha)} - \sum_{\beta} |a_\beta| \sum_{x_\beta} p_\beta(x_\beta) \log p_\beta(x_\beta) + \sum_{\gamma} a_\gamma \sum_{x_\gamma} p_\gamma(x_\gamma) \log p_\gamma(x_\gamma)$$

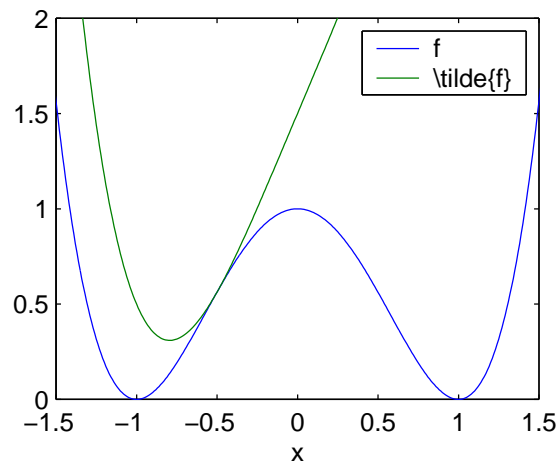
subject to $\sum_{x_\alpha} p_\alpha(x_\alpha) = 1, p_\alpha(x_\beta) = p_\beta(x_\beta), \beta \subset \alpha, p_\alpha(x_\alpha) \geq 0.$



Double loop approach

Bound $f(x)$ by a convex function:

$$f(x) \leq \tilde{f}_{x_0}(x) \quad f(x_0) = \tilde{f}_{x_0}(x_0)$$

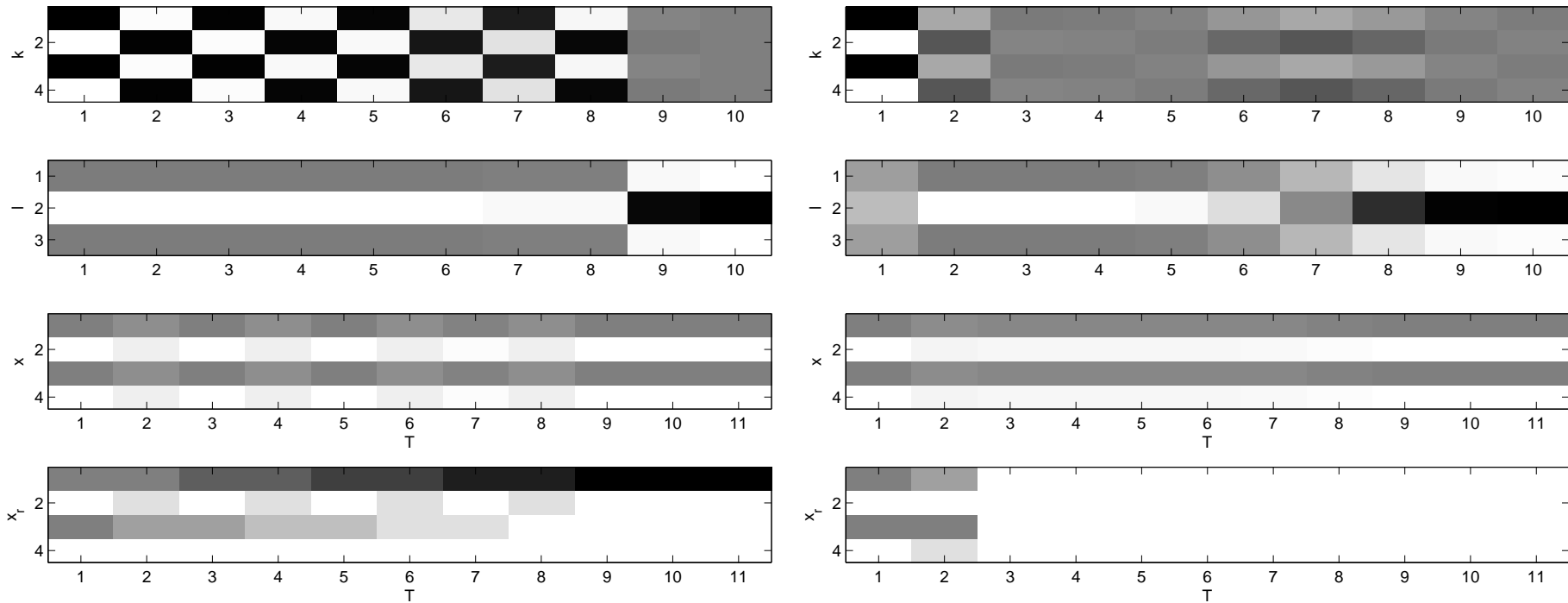


Then, optimizing $\tilde{f}_{x_0}(x)$ wrt x under constraints is a convex problem that can be solved, and

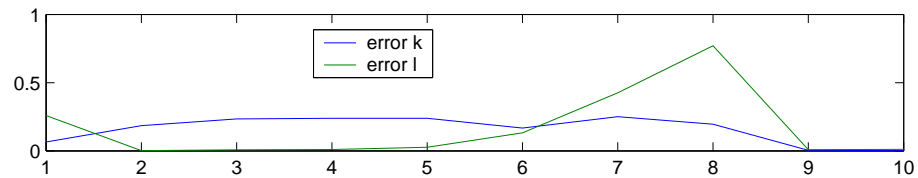
$$f(x_0) = \tilde{f}_{x_0}(x_0) \geq \tilde{f}_{x_0}(x^*(x_0)) \geq f(x^*(x_0))$$

Heskes, Albers, Kappen, UAI 2003

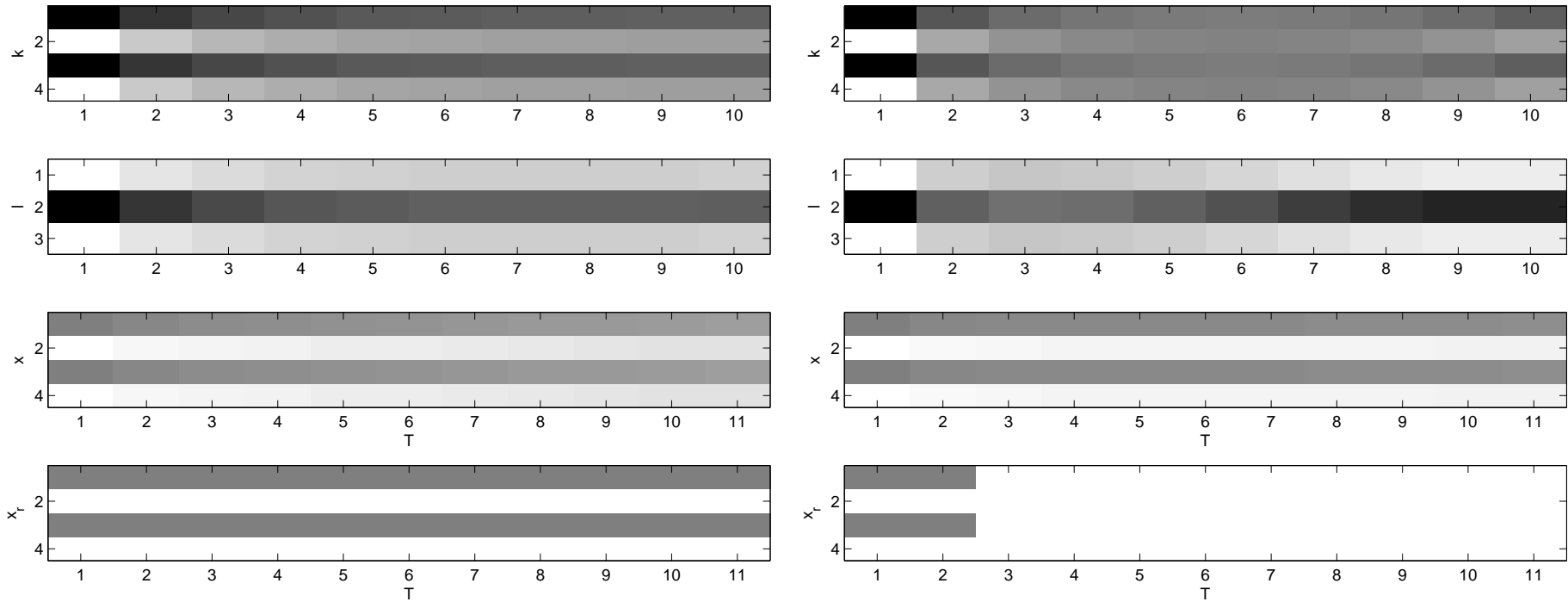
A blocks world



$n = 4, m = 8, T = 10, \lambda = 10$ using exact inference (left) and CVM (right).



A blocks world



$n = 4, m = 8, T = 10, \lambda = 2$ using exact inference (left) and CVM (right).

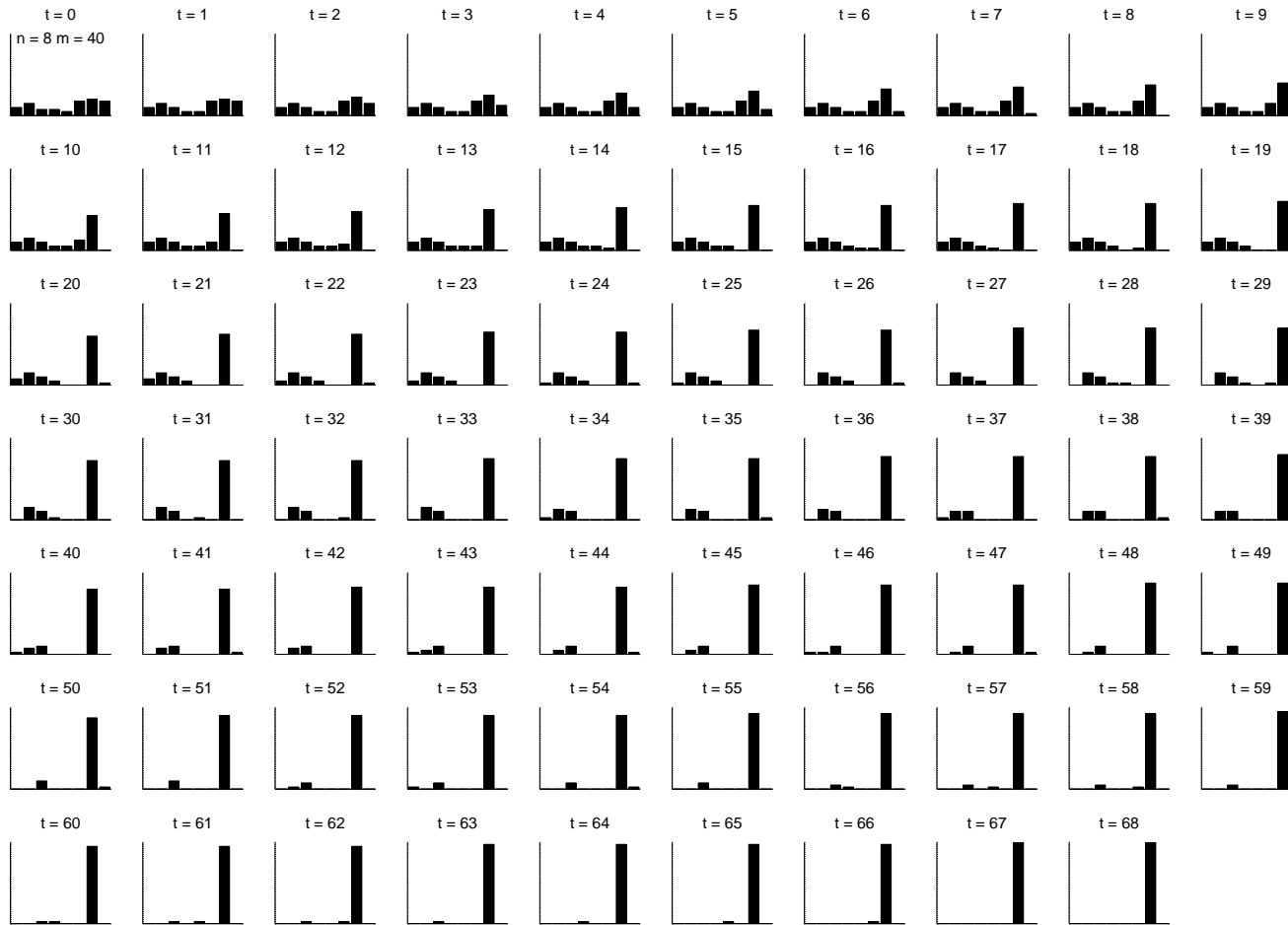
A blocks world

The computational requirements for exact computation scale very fast with n and m .

n	m	T	cliquesize	exact (Mb)	CVM (Mb)	Max error	Max error $T=1$
4	2	11	7	17	2.2	0.0304	0.0158
4	4	11	7	132	2.3	0.2348	0.1066
6	2	11	12	680	2.7	0.9596	0.0174
6	4	11	12	15.000	2.9	0.9732	0.1265

Table 1: CVM errors for some small examples. Horizon time $T = 10$. Initial block configuration is symmetric with $m/2$ blocks on two stacks maximally separated. CVM was used with 50 innerloops and a stop crit=0.00001.

A larger example



$n = 8, m = 40, T = 80, \lambda = 10$ using CVM. CPU time per iteration approx 2000-4000 sec.
memory use 27 Mb.

Summary and discussion

KL control problems:

control cost is KL divergence between future trajectories
contains diffusion processes and jump processes as special cases

Exploit graphical structure in optimal control computation:

exact inference
approximate inference

Agents

Partial observability

Agents

Assume that

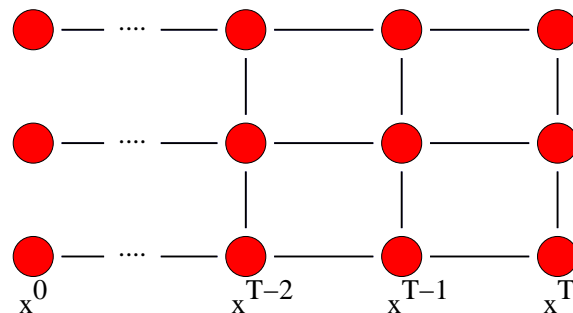
q factorizes: $q_{xy} = \prod_{i=1}^n q_i(y_i|x_i)$.

p factorizes: $p_{xy} = \prod_{i=1}^n p_i(y_i|x_i)$.

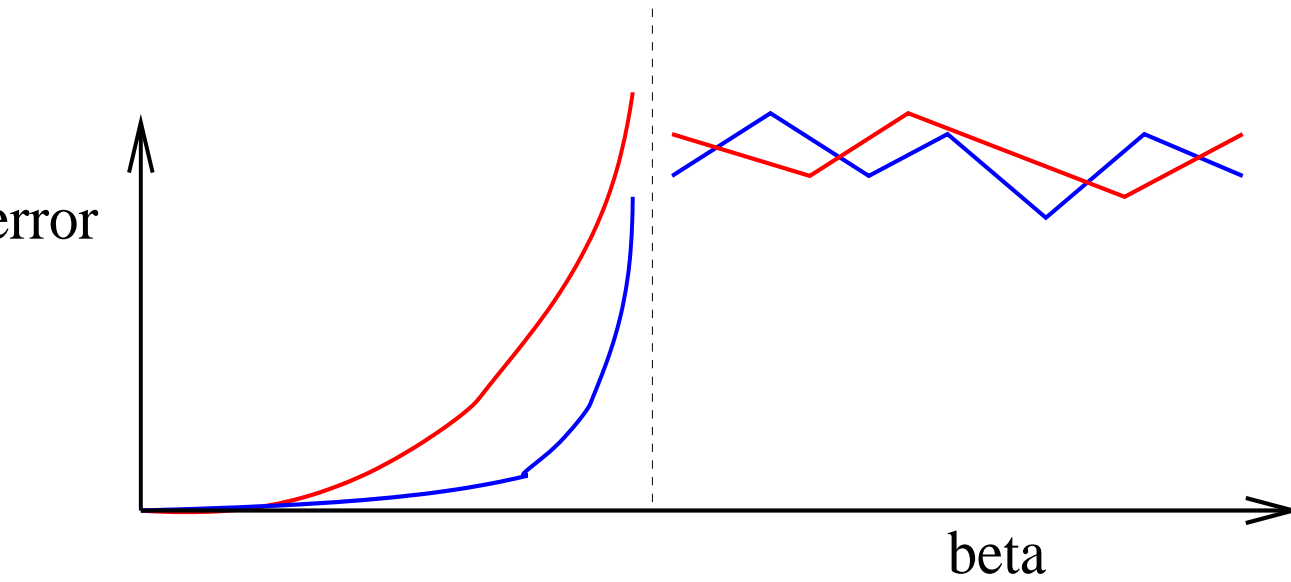
R has a sparse structure: $R(x) = \sum_{ij} R_{ij}(x_i, x_j)$.

MF ansatz is 'natural' rather than an approximation.

$$C = \sum_i KL(p_i||q_i) + \sum_{ij} \langle R_{ij} \rangle_p$$

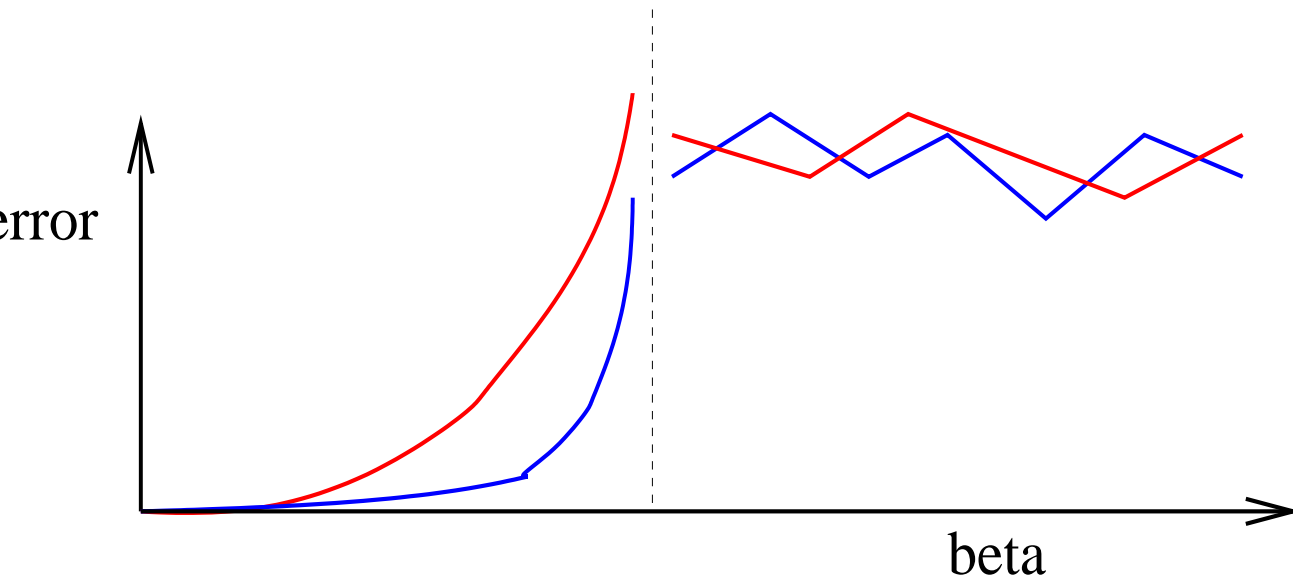


Where do we stand...



- Variational
- MCMC
- BP/EP
- GBP/CVM
- TRW
- Bounds
- Loop corrections

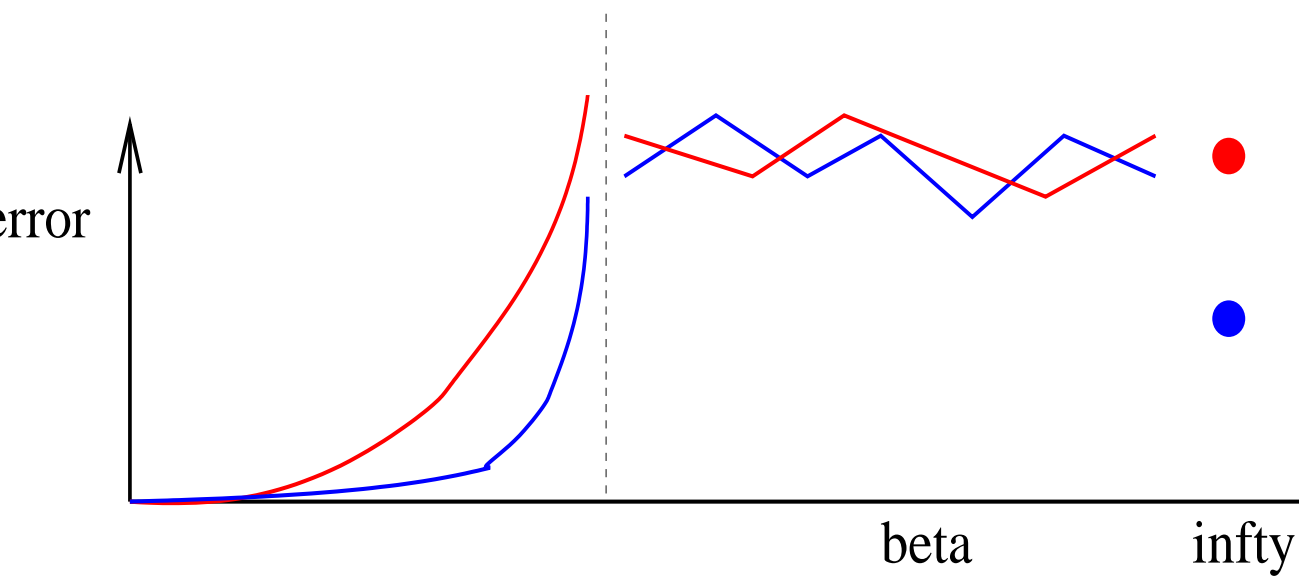
Where do we stand...



- Variational
- MCMC
- BP/EP
- GBP/CVM
- TRW
- Bounds
- Loop corrections

Saturation has been reached

Where do we stand...



Variational
MCMC
BP/EP
GBP/CVM
TRW
Bounds
Loop corrections

Convex optimization
Combinatoric optimization

Saturation has been reached

Path integral

We previously established

$$p(x^{1:T}|x^0) = \frac{1}{Z(x^0)} r(x^{1:T}|x^0)$$

$$\beta^t(x^t) = \sum_{x^{t+1:T}} r(x^{t+1:T}|x^t) = \sum_{x^{t+1:T}} q(x^{t+1:T}|x^t) \exp\left(-\sum_{s=t}^T R(x^s, s)\right)$$

Thus,

$$-\log \beta^0(x^0) = -\log \sum_{x^{1:T}} r(x^{1:T}|x^0) = -\log Z(x^0) = C_{\min}$$