Control, inference and learning

Bert Kappen
: SNN Donders Institute, Radboud University, Nijmegen
Gatsby Unit, UCL London

July 21, 2015

Why control theory?

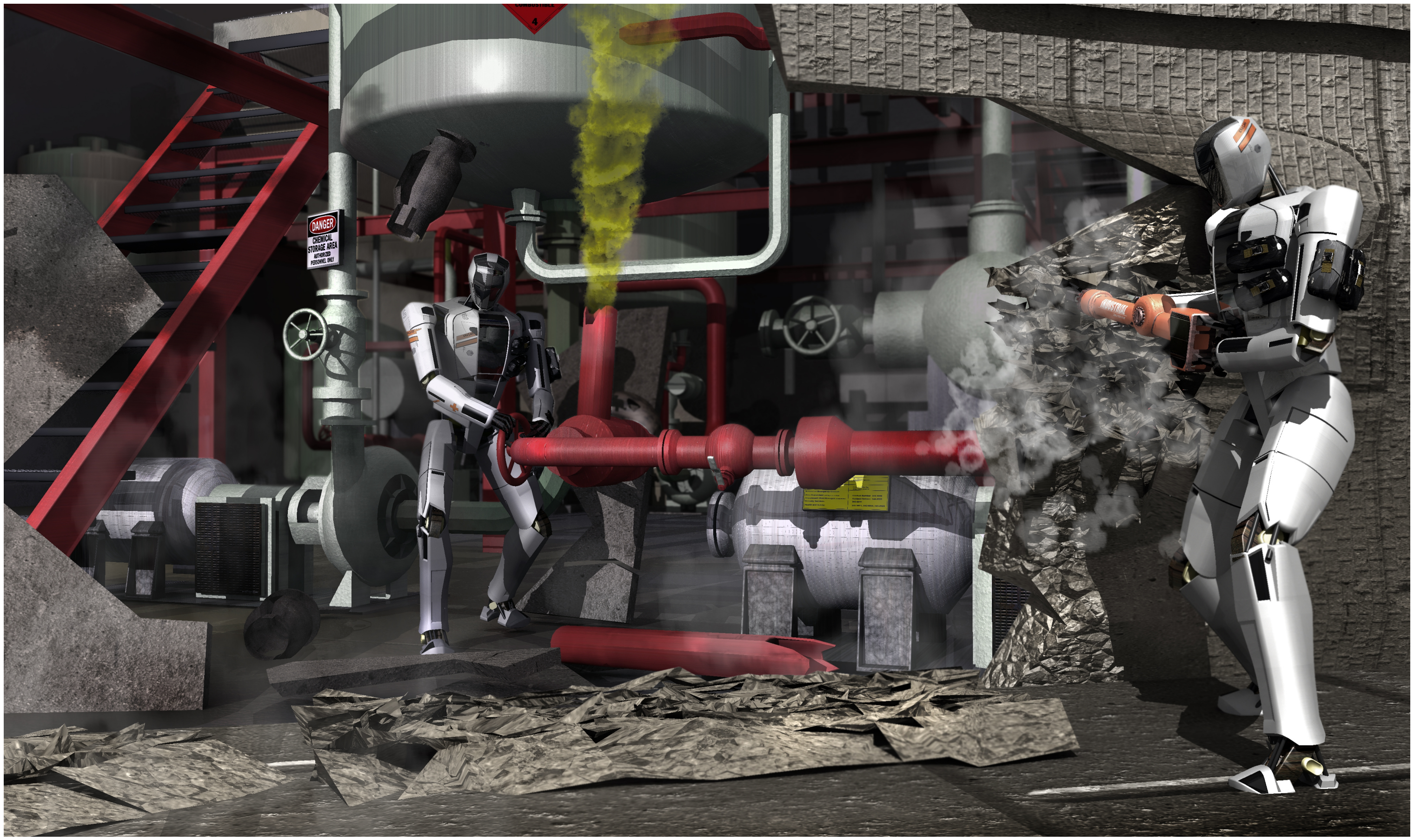
A theory for intelligent behaviour:

- neuroscience

Why control theory?

A theory for intelligent behaviour:

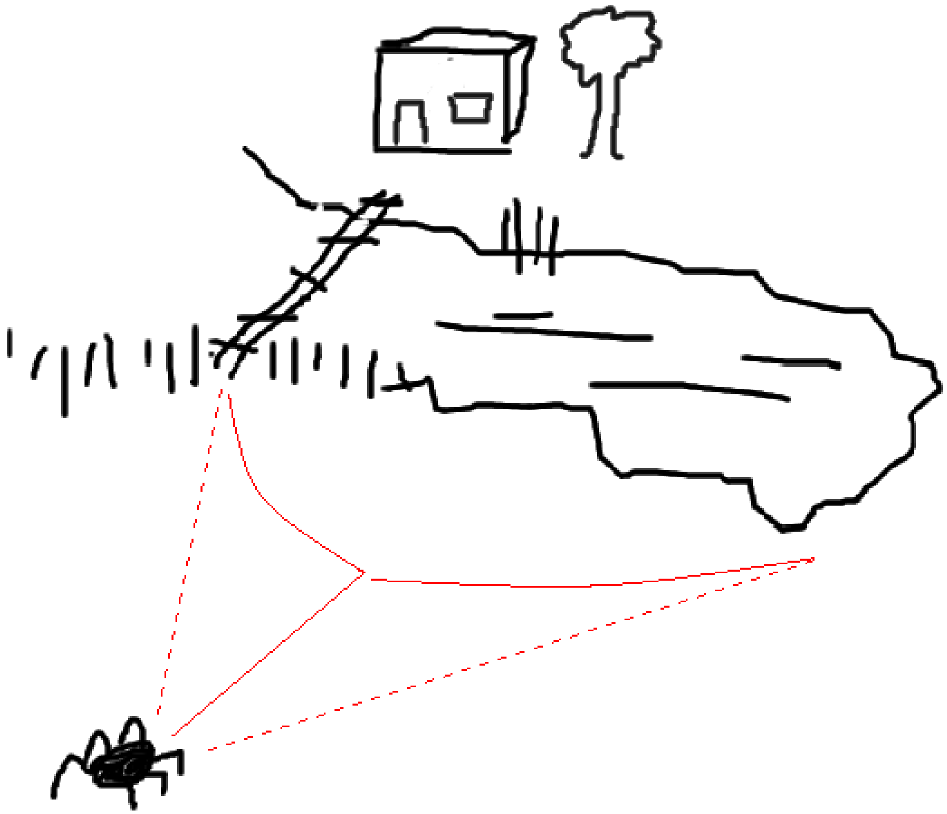
- neuroscience
- robotics



Control theory

Given a current state and a future desired state, what is the best/cheapest/fastest way to get there.

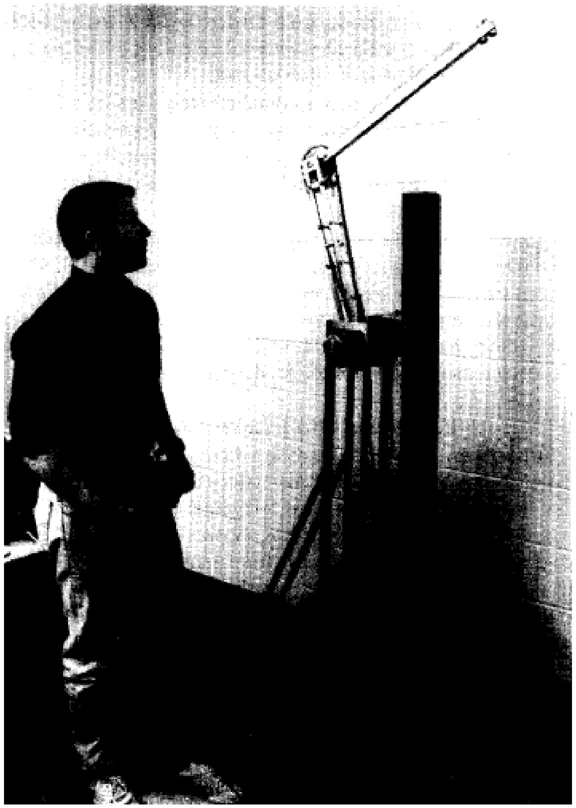
Why stochastic control?



How to control?

Hard problems:

- a learning and exploration problem
- a stochastic optimal control computation
- a representation problem $u(x, t)$



The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution
Express a control computation as an inference computation.

The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution
Express a control computation as an inference computation.
Compute optimal control using MC sampling

The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution
Express a control computation as an inference computation.
Compute optimal control using MC sampling

Importance sampling
Accellerate with importance sampling, a state-feedback controller

The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution

Express a control computation as an inference computation.
Compute optimal control using MC sampling

Importance sampling

Accellerate with importance sampling, a state-feedback controller
Learn controller from self-generated data

The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution
Express a control computation as an inference computation.
Compute optimal control using MC sampling

Importance sampling
Accellerate with importance sampling, a state-feedback controller
Learn controller from self-generated data

Optimal importance sampler is optimal control

The idea: Control, Inference and Learning

Linear Bellman equation and path integral solution
Express a control computation as an inference computation.
Compute optimal control using MC sampling

Importance sampling
Accellerate with importance sampling, a state-feedback controller
Learn controller from self-generated data

Optimal importance sampler is optimal control

Learn a good importance sampler using PICE

Outline

- Introduction to control theory
- Link between control theory, inference and statistical physics
  - Schrödinger, Fleming Mitter ’82, Kappen ’05, Todorov ’06
- Importance sampling
  - Relation between optimal sampling and optimal control
- Cross entropy method for adaptive importance sampling (PICE)
  - A criterion for parametrized control optimization
  - Learning by gradient descent
- Some examples

Discrete time optimal control

Consider the control of a discrete time deterministic dynamical system:

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

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

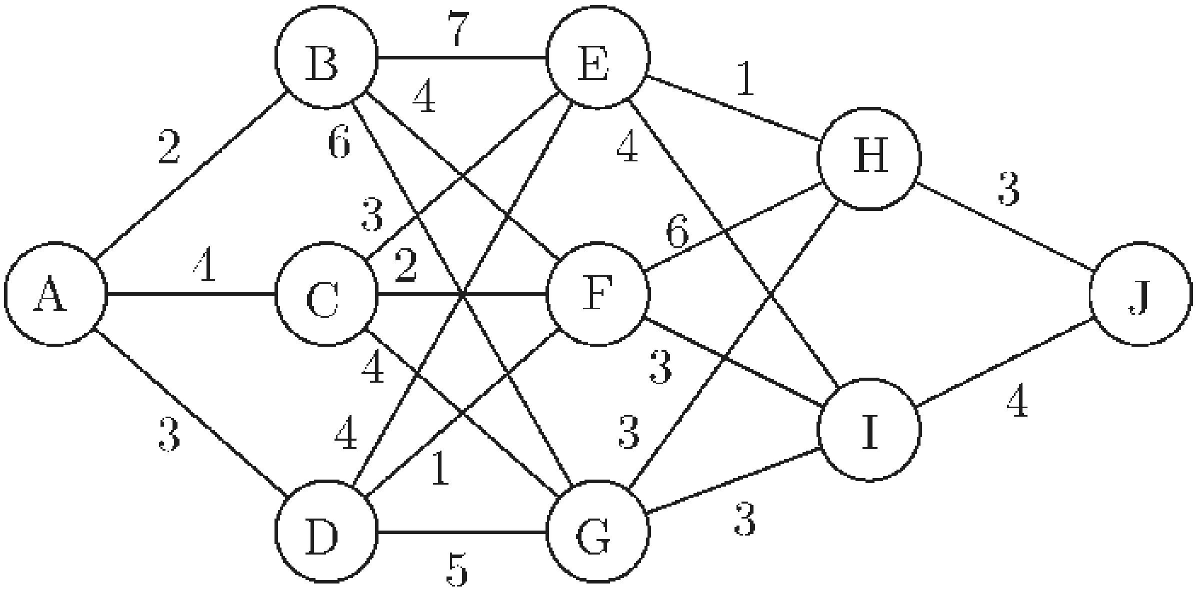
Given x_0 and $u_{0:T-1}$, we can compute $x_{1:T}$.

Define a cost for each sequence of controls:

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

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.

$$C(F) \quad = \quad \min(6 + C(H), 3 + C(I)) = 7$$

Minimal cost at time t easily expressable in terms of minimal cost at time $t + 1$.

Discrete time optimal control

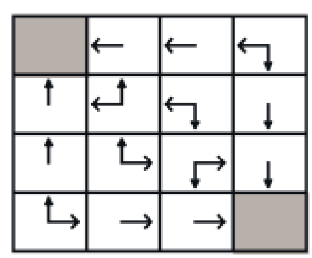
Dynamic programming uses concept of optimal cost-to-go $J(t, x)$.

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

$$
\begin{aligned}
J(t, x_t) &= \min_{u_{t:T-1}} \left(\sum_{s=t}^{T-1} V(x_s, u_s) \right) \\
&= \min_{u_t} \left(V(t, x_t, u_t) + J(t+1, x_t + f(t, x_t, u_t)) \right) \\
J(T, x) &= 0 \\
J(0, x) &= \min_{u_{0:T-1}} C(x, u_{0:T-1})
\end{aligned}
$$

This is called the Bellman Equation. Computes $u_t(x)$ for all intermediate t, x.

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0



Stochastic optimal control

Consider a stochastic dynamical system

$$dX_i = f_i(X_t, u)dt + dW_i \qquad \mathbb{E}(dW_i dW_j) = \nu_{ij} dt$$

Given $x(0)$ find control function $u(x,t)$ that minimizes the expected future cost

$$C = \mathbb{E}\left(\phi(X_T) + \int_0^T dt V(X_t, u(X_t, t))\right)$$

Expectation is over all trajectories given the control path.

$$\begin{aligned} J(t,x) &= \min_u \left(V(x,u) + \mathbb{E}\, J(t+dt, x+dx)\right) \\ -\partial_t J(t,x) &= \min_u \left(V(x,u) + f(x,u)\nabla_x J(x,t) + \frac{1}{2}\nu \nabla_x^2 J(x,t)\right) \end{aligned}$$

with $u = u(x,t)$ and boundary condition $J(x,T) = \phi(x)$. This is HJB equation.

Computing the optimal control solution is hard

- solve a Bellman Equation, a PDE
- scales badly with dimension

Efficient solutions exist for

- linear dynamical systems with quadratic costs (Gaussians)
- deterministic systems (no noise)

Path integral control theory

$$
\begin{aligned}
dX_t &= f(X_t,t)dt + g(X_t,t)(u dt + dW_t) \\
C &= \mathbb{E}\left(\phi(X_T) + \int_t^T ds V(X_s,s) + \frac{1}{2}u^T(X_t,t)Ru(X_t,t)\right)
\end{aligned}
$$

with $\mathbb{E}(dW_a dW_b) = \nu_{ab}dt$ and $R = \lambda\nu^{-1}, \lambda > 0$. $f \in \mathbb{R}^n, g \in \mathbb{R}^{n\times m}, u \in \mathbb{R}^m$.

The HJB equation becomes

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

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

Path integral control theory

Minimization wrt u yields:

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

Define $\psi(x,t)$ through $J(x,t) = -\lambda \log \psi(x,t)$. We obtain a linear HJB:

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

Feynman-Kac formula

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

$$dX_t \quad = \quad f(X_t,t)dt + g(X_t,t)dW$$

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

$$\begin{aligned} \psi(x,t) &= \int dq(\tau|x,t)\exp\left(-\frac{S(\tau)}{\lambda}\right) = \mathbb{E}_q\left(e^{-S/\lambda}\right) \\ S(\tau) &= \phi(x(T)) + \int_t^T ds V(x(s),s) \end{aligned}$$

Posterior distribution over optimal trajectories

$\psi(x,t)$ is the 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:

$$J(x,t) = -\lambda \log \mathbb{E}_q \left(e^{-S/\lambda}\right)$$

The optimal control is an expectation wrt p:

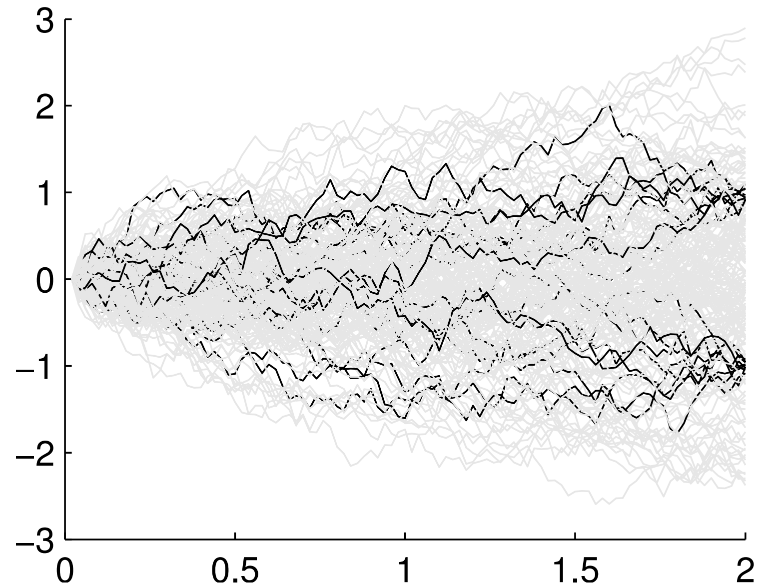
$$u^*(x,t)dt = \mathbb{E}_{p^*}(dW_t) = \frac{\mathbb{E}_q \left(dW e^{-S/\lambda}\right)}{\mathbb{E}_q \left(e^{-S/\lambda}\right)}$$

J, u^* can be computed by forward sampling from q.

Delayed choice

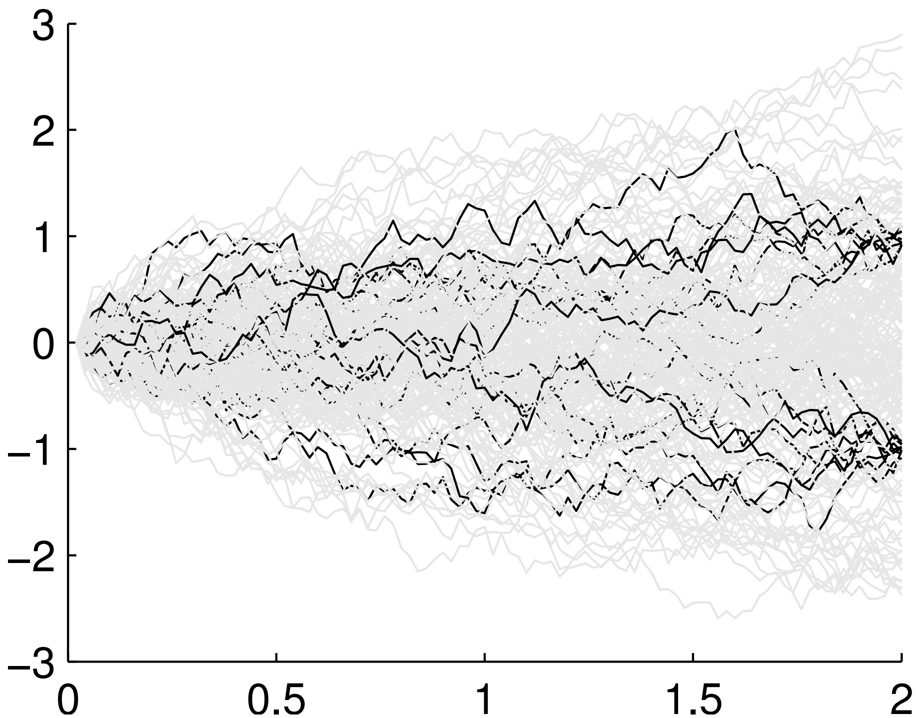
$$
\begin{aligned}
dX_t &= u(X_t, t)dt + dW_t \qquad \left\langle dW_t^2 \right\rangle = \nu dt \\
C(p) &= \mathbb{E}_p \phi(x_T) + \int_0^2 dt \frac{1}{2} u(t)^2
\end{aligned}
$$

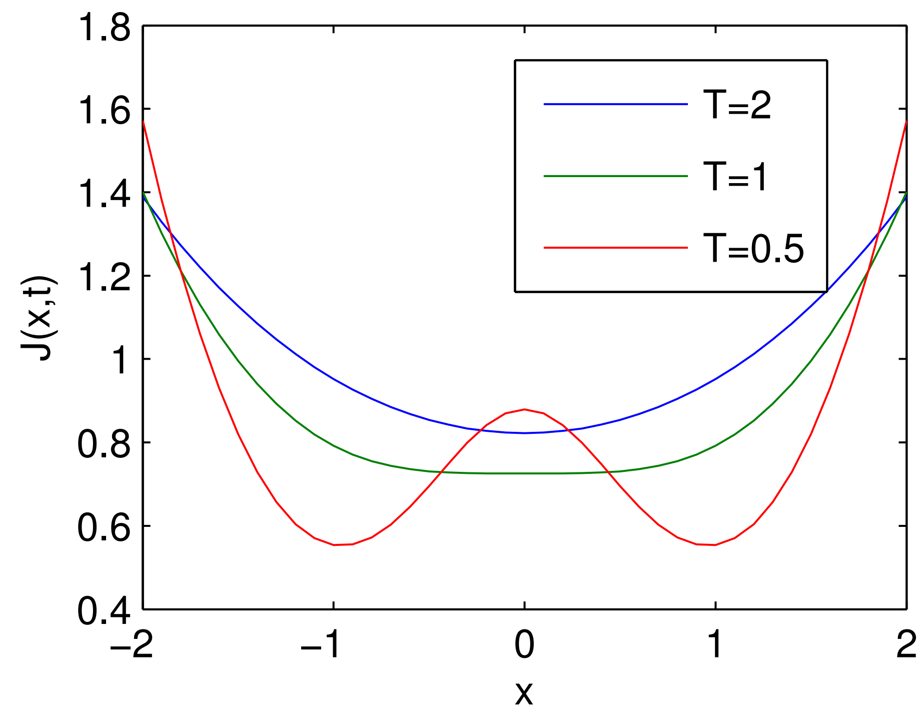
Cost encodes targets at $t = 2$.



Delayed choice

Time-to-go $T = 2 - t$.



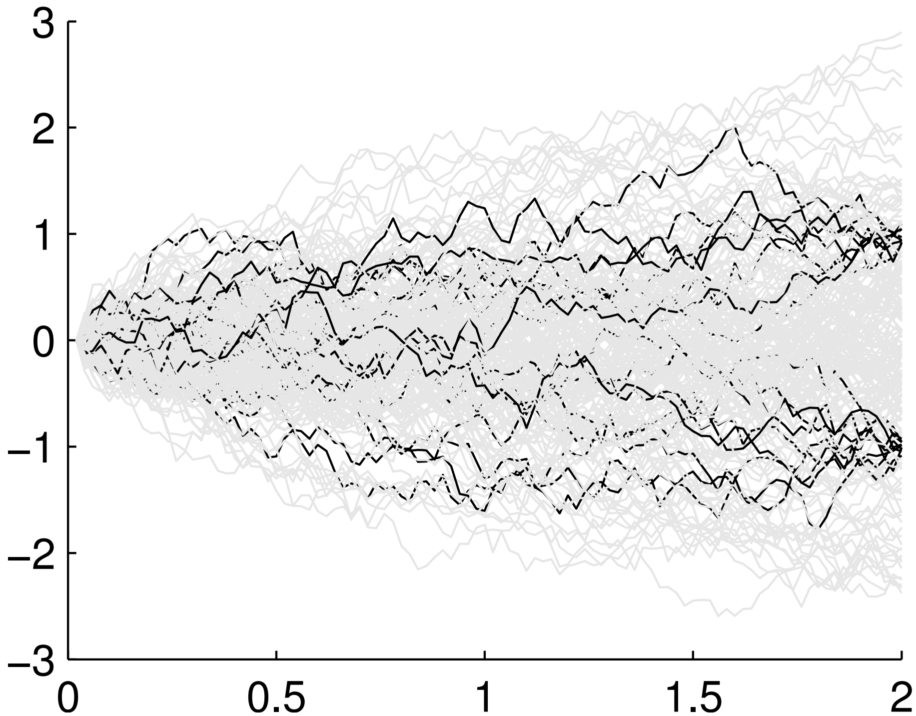


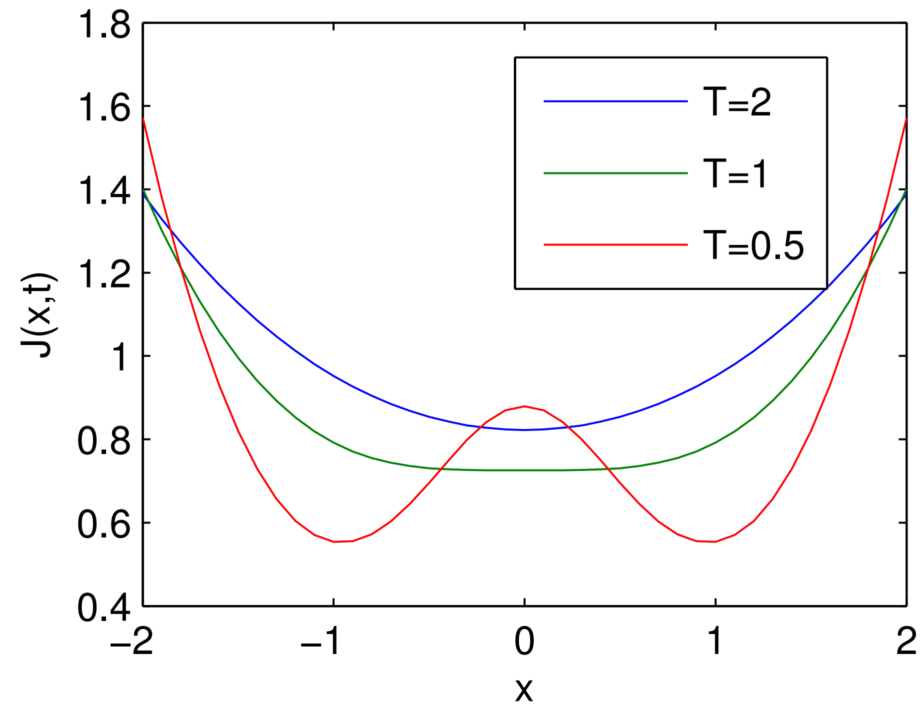
$$J(x, t) = -\nu \log \mathbb{E}_q \exp(-\phi(X_2)/\nu)$$

Decision is made at $T = \frac{1}{\nu}$

Delayed choice

Time-to-go $T = 2 - t$.

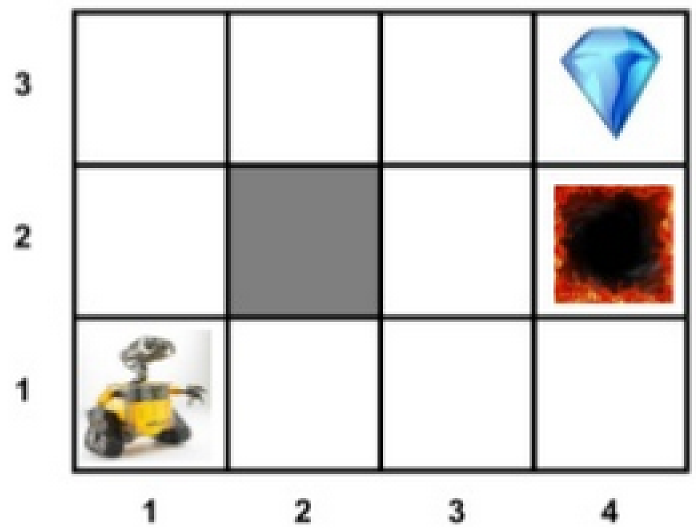


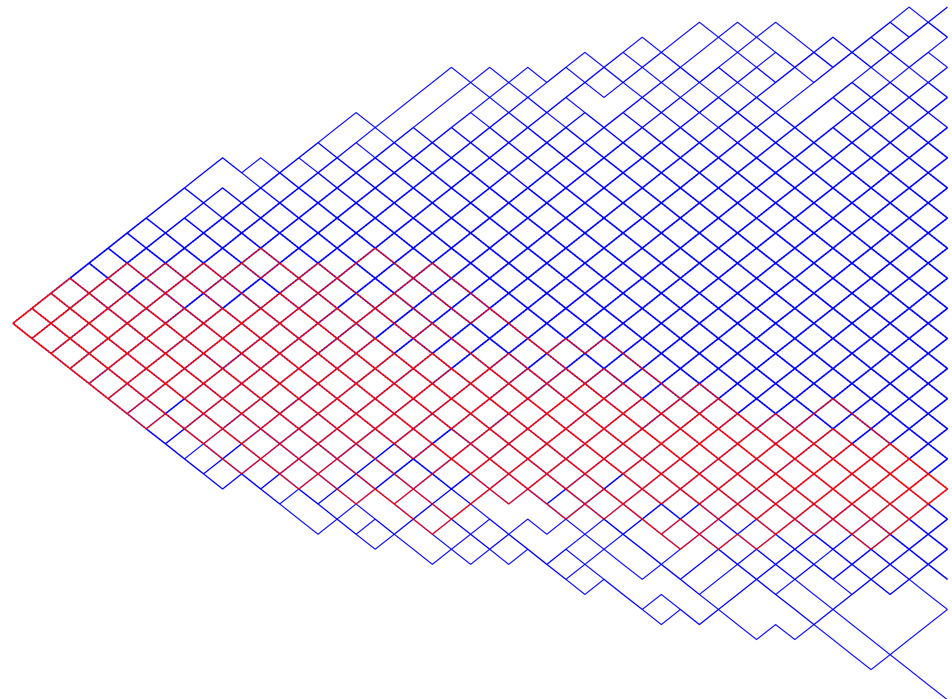


$$J(x,t) = -\nu \log \mathbb{E}_q \exp(-\phi(X_2)/\nu)$$

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

KL control





Uncontrolled dynamics specifies distribution $q(\tau|x,t)$ over trajectories τ from $t \to T$.

Cost for trajectory τ is $S(\tau) = \phi(x_T) + \int_t^T ds V(x_s, s)$.

Find optimal distribution $p(\tau|x.t)$ that minimizes $\mathbb{E}_p\, S$ and is 'close' to $q(\tau|x,t)$.

KL control

Find p^* that minimizes

$$C(p) = KL(p|q) + \mathbb{E}_p\, S \qquad KL(p|q) = \int d\tau p(\tau|x,t) \log \frac{p(\tau|x,t)}{q(\tau|x,t)}$$

The optimal solution is given by

$$p^*(\tau|x,t) = \frac{1}{\psi(x,t)} q(\tau|x,t) \exp(-S(\tau|x,t)) \qquad \psi(x,t) = \int d\tau q(\tau|x,t) \exp(-S(\tau|x,t))$$

The optimal cost is:

$$C(p^*) = -\log \psi(x,t)$$

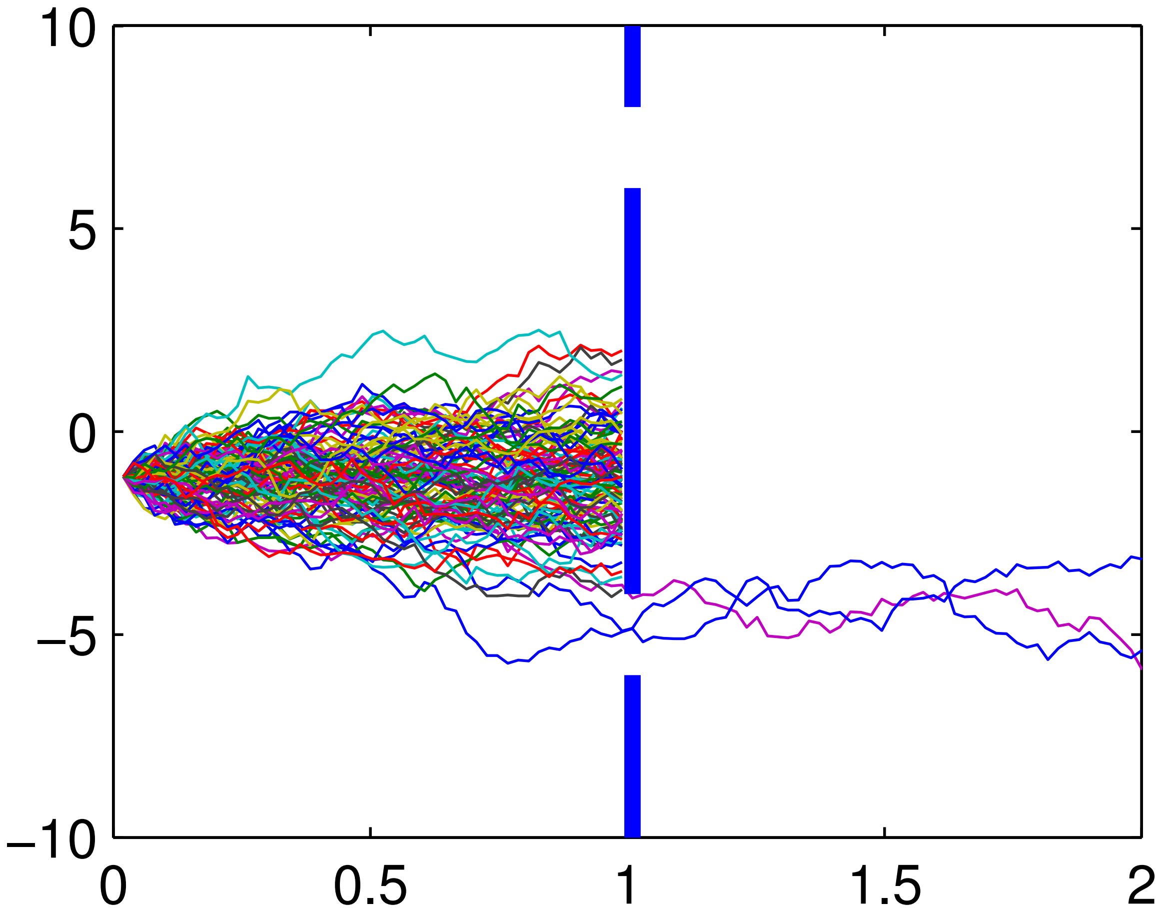
Controlled diffusions are special case

In the case of controlled diffusions, p is parametrised by functions $u(x,t)$:

$$\begin{aligned} dX_t &= f(X_t,t)dt + g(X_t,t)(u(X_t,t)dt + dW_t) \qquad \mathbb{E}(dW_i dW_j) = \nu_{ij}dt \\ C(p) &= \mathbb{E}_p\left(\phi(X_T) + \int_t^T ds \frac{1}{2} u(X_s,s)^T \nu^{-1} u(X_s,s) + V(X_s,s)\right) \end{aligned}$$

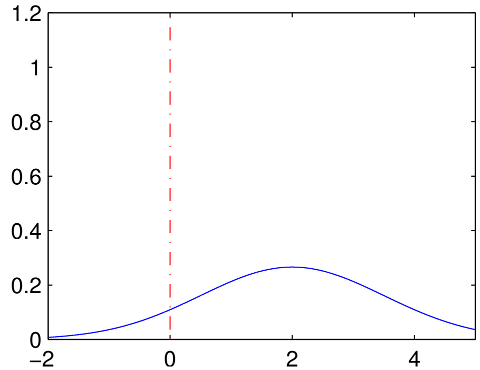
$\psi(x,t)$ is the solution of the linear Bellman equation and $J(x,t) = -\log\psi(x,t)$ is the optimal cost-to-go.

Sampling efficiency



Sampling with uncontrolled dynamics is theoretically correct, but inefficient in efficient in practice.

Importance sampling



Consider simple 1-d sampling problem. Given $q(x)$, compute

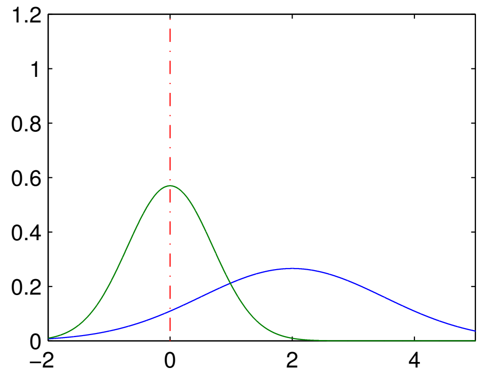
$$a = \text{Prob}(x < 0) = \int_{-\infty}^{\infty} I(x) q(x) dx$$

with $I(x) = 0, 1$ if $x > 0, x < 0$, respectively.

Naive method: generate N samples $X_i \sim q$

$$\hat{a} = \frac{1}{N} \sum_{i=1}^{N} I(X_i)$$

Importance sampling



Consider another distribution $p(x)$. Then

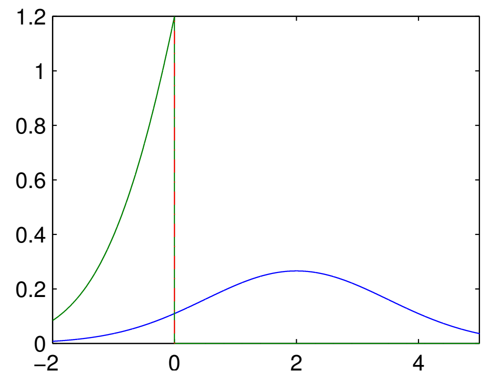
$$a = \text{Prob}(x < 0) = \int_{-\infty}^{\infty} I(x) \frac{q(x)}{p(x)} p(x) dx$$

Importance sampling: generate N samples $X_i \sim p$

$$\hat{a} = \frac{1}{N} \sum_{i=1}^{N} I(X_i) \frac{q(X_i)}{p(X_i)}$$

Unbiased (= correct) for any p!

Optimal importance sampling



The distribution

$$p^*(x) = \frac{q(x)I(x)}{a}$$

is the optimal importance sampler. One sample $X_i \sim p^*$ is sufficient to estimate a:

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} I(X_i)\frac{q(X_i)}{p^*(X_i)} = a$$

"Optimal importance sampler has zero variance".

Importance sampling and control

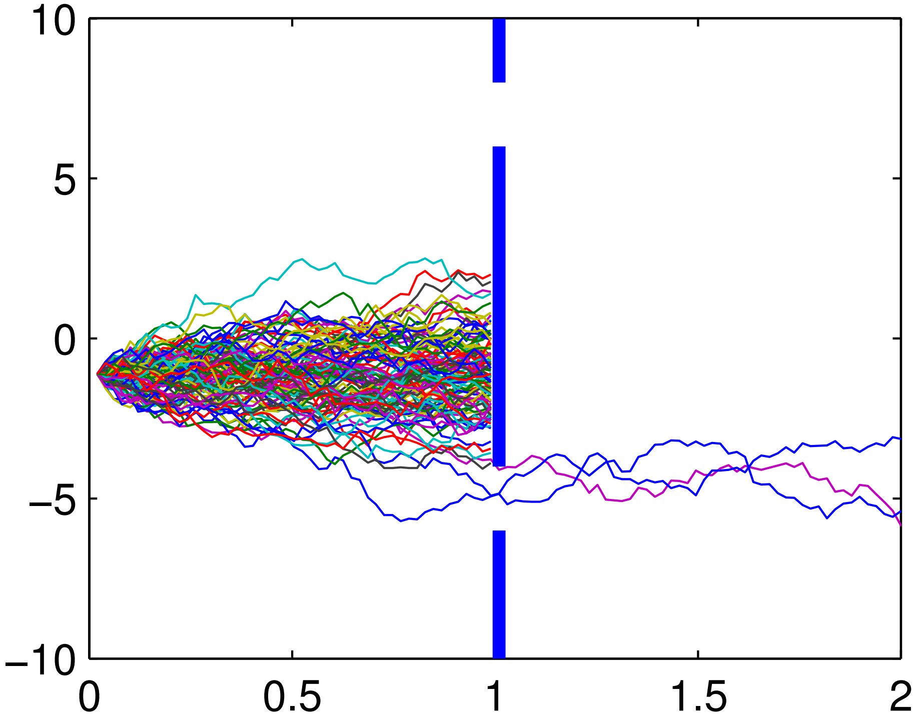
Theorem 1. *The solution of the control problem is given by*

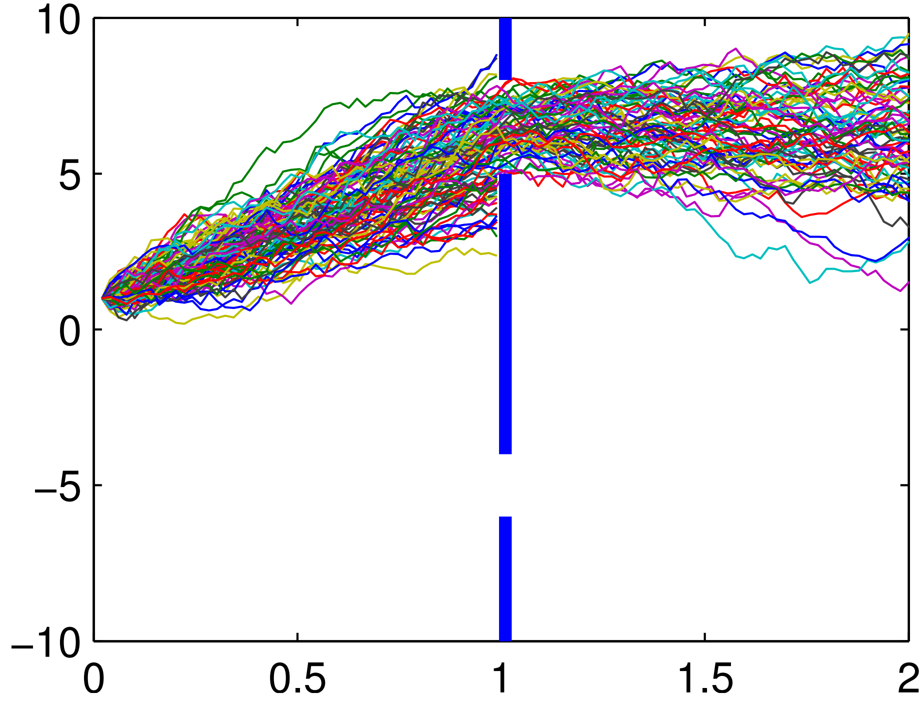
$$
\begin{aligned}
J(x,t) &= -\log E_q e^{-S} = -\log \mathbb{E}_p e^{-S} \frac{dq}{dp} = -\log \mathbb{E}_u e^{-S^u} \\
u^*(x,t)dt &= \frac{\mathbb{E}_q\left(dW_t e^{-S}\right)}{\mathbb{E}_q\left(e^{-S}\right)} = u(t,x)dt + \frac{\mathbb{E}_u\left(dW_t e^{-S^u}\right)}{\mathbb{E}_u\left(e^{-S^u}\right)} \\
\frac{dq}{dp} &= \exp\left(-\int_t^T dt \frac{1}{2} u(x,t)^T \nu^{-1} u(x,t) - \int_t^T u(x,t)^T \nu^{-1} dW_t\right)
\end{aligned}
$$

with $\mathbb{E}_p = \mathbb{E}_u$.

We can choose any p, ie. any sampling control u.

Importance sampling and control





Relation between optimal sampling and optimal control

Definition 2.

1. *The weight of a path is defined as* $\alpha^u = \frac{e^{-S^u(t_0,x_0)}}{\mathbb{E}[e^{-S^u(t_0,x_0)}]}$.
2. *The fraction of effective samples is* $FES = \frac{1}{\mathbb{E}[(\alpha^u)^2]} = \frac{1}{\mathrm{Var}(\alpha^u)+1}$.

Theorem 3. *Let* $0 < \epsilon < 1$. *Then:*

1. $(u^* - u)'(u^* - u) \leq \frac{\epsilon}{t_1 - t_0}$ *point-wise implies* $\mathrm{Var}\,(\alpha^u) \leq \frac{\epsilon}{1-\epsilon}$
2. $\mathrm{Var}(\alpha^u) \leq \epsilon$ *implies* $\int_{t_0}^{t_1} \langle u^* - u \rangle' \langle u^* - u \rangle \, dt \leq \epsilon$.

1. Better u (in the sense of optimal control) provides a better sampler (in the sense of effective sample size).
2. Optimal $u = u^*$ (in the sense of optimal control) requires only one sample.

The Cross-entropy method

Let X be a random variable taking values in the space $\mathcal{X}$. Let $f_v(x)$ be a family of probability density function on $\mathcal{X}$ parametrized by v and $h(x)$ be a positive function. Suppose that we are interested in the expectation value

$$a = \mathbb{E}_u\, h = \int dx f_u(x) h(x)$$

where $\mathbb{E}_u$ denotes expectation with respect to the pdf f_u for a particular value of $v = u$.

The optimal importance sampling distribution is $g^*(x) = h(x) f_u(x)/a$.

The cross entropy method suggests to find the distribution f_v in the parametrized family of distributions that minimises the KL divergence

$$KL(g^*|f_v) = \int dx g^*(x) \log \frac{g^*(x)}{f_v(x)} \propto -\mathbb{E}_{g^*} \log f_v(X) \propto -\mathbb{E}_u h(X) \log f_v(X) = -D(v)$$

The Cross-entropy method

We can use again importance sampling to compute $D(v)$:

$$D(v) = \mathbb{E}_u h(X) \log f_v(X) = \mathbb{E}_w h(X) \frac{f_u(X)}{f_w(X)} \log f_v(X)$$

We estimate the expectation value by drawing N samples from f_w. If D is convex and differentiable with respect to v, the optimal v is given by

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \frac{f_u(X_i)}{f_w(X_i)} \frac{d}{dv} \log f_v(X_i) = 0 \qquad X_i \sim f_w$$

The CE algorithm

Initialize $w_0 = u$.
for $k = 0, \ldots, K$ **do**
generate N samples $X_{1:N}$ from f_{w_k}
compute v by solving

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i)\frac{f_u(X_i)}{f_w(X_i)}\frac{d}{dv}\log f_v(X_i) = 0$$

Set $w_{k+1} = v$.
end for
return w_K

The CE method for PI control: Preliminaries

Let $\mathcal{X}$ denote the space of continuous trajectories on the interval $[t, T]$: $\tau = X(s), t \leq s \leq T$ with fixed initial value $X(t) = x$ satisfying the dynamics

$$dX_t = f(X_t, t)dt + g(X_t, t)\left(u(X_t, t)dt + dW_t\right)$$

Denote $p_u(\tau)$ the distribution over trajectories τ with control u.

The distributions p_u and p_0 are related by the Girsanov Theorem.

$$\begin{aligned}
p(X_{s+ds}|X_s) &= \mathcal{N}(X_{s+ds}|\mu_s, \Xi_s ds) \qquad \mu_s = X_s + \mathbb{E}dX_s \qquad \Xi_s = \mathbb{E}dX_s^2 \\
p_u(\tau) &= \lim_{ds \to 0} \prod_{s=t}^{T-ds} \mathcal{N}(X_{s+ds}|\mu_s, \Xi_s) \\
&= p_0(\tau) \exp\left(-\int_t^T ds \frac{1}{2} u^2(s, X_s) + \int_t^T u(s, X_s) g(s, X_s)^{-1} (dX_s - f(s, X_s) ds)\right)
\end{aligned}$$

The Radon-Nikodym can be used to rewrite the optimal distribution:

$$
\begin{aligned}
\frac{dp_0(\tau)}{dp_u(\tau)} &= \exp\left(-\int_t^T ds \frac{1}{2} u^2(s, X(s)) - \int_t^T u(s, X(s)) dW(s)\right) \\
p^*(\tau) &= \frac{1}{\psi(t,x)} p_0(\tau) \exp(-V(\tau)) = \frac{1}{\psi(t,x)} p_u(\tau) \frac{dp_0(\tau)}{dp_u(\tau)} \exp(-V(\tau)) \\
&= \frac{1}{\psi(t,x)} p_u(\tau) \exp(-S(t,x,u))
\end{aligned}
$$

The CE method for PI control

We have a family of distributions p_u. We wish to compute a near optimal control $\hat{u}$ such that $p_{\hat{u}}$ is close to p^*. Following the CE argument, we minimise

$$
\begin{aligned}
KL(p^*|p_{\hat{u}}) &= \mathbb{E}_{p^*} \log p^* - \mathbb{E}_{p^*} \log p_{\hat{u}} \propto -\mathbb{E}_{p^*} \log p_{\hat{u}} \\
&\propto \mathbb{E}_{p^*} \left(\int_t^T \frac{1}{2} \hat{u}^2(s, X_s) ds - \hat{u}(s, X_s) g(s, X_s)^{-1} (dX_s - f(s, X_s) ds) \right) \\
&= \frac{1}{\psi(t, x)} \mathbb{E}_p e^{-S(t,x,u)} \int_t^T ds \left(\frac{1}{2} \hat{u}(s, X(s))^2 - \hat{u}(s, X(s)) \left(u(s, X(s)) + \frac{dW_s}{ds} \right) \right)
\end{aligned}
$$

The expression must be optimized with respect to the functions $\hat{u}_{t:T} = \{\hat{u}(s, X_s), t \leq s \leq T\}$. It is independent of the sampling control $u_{t:T} = \{u(s, X_s), t \leq s \leq T\}$.

The CE method for PI control: Time-dependent solution

We now assume that $\hat{u}$ is a parametrized function with parameters θ. In the time-dependent case, we consider different θ_s for each of the functions $\hat{u}(s, x|\theta_s)$ separately. The gradient is given by:

$$\frac{\partial KL(p^*|\hat{p})}{\partial \theta_s} = \frac{1}{\psi(t,x)} \mathbb{E}_p e^{-S(t,x,u)} \left(\hat{u}(s, X(s)) - u(s, X(s)) - \frac{dW_s}{ds} \right) \frac{\partial \hat{u}(s, X(s))}{\partial \theta_s}$$

Choosing $u = \hat{u}$ yields the gradient procedure

$$\theta_{s,n+1} = \theta_{s,n} - \eta \frac{\partial KL(p^*|\hat{p})}{\partial \theta_{s,n}}\Big|_{u=\hat{u}_n} = \theta_{s,n} + \eta \left\langle \frac{dW_s}{ds} \frac{\partial \hat{u}(s, X(s))}{\partial \theta_{s,n}} \right\rangle$$

with $\langle F \rangle = \frac{1}{\psi(t,x)} \mathbb{E}_p e^{-S(t,x,u)} F$ and $\eta > 0$ a small parameter.

Convergence is guaranteed. We refer to this gradient method as PICE.

The CE method for PI control: Time-dependent solution

Linear basis functions:

$$\hat{u}(s,x) = \sum_{k=1}^{K} \theta_{sk} h_{sk}(x) \qquad u(s,x) = \sum_{k=1}^{K} \theta^0_{sk} h_{sk}(x)$$

we obtain regression problem:

$$\sum_{l=1}^{K} \left(\theta_{sl} - \theta^0_{sl}\right) \langle h_{sl} h_{sk} \rangle = \left\langle \frac{dW_s}{ds} h_{sk} \right\rangle$$

For each s a system of K linear equations with K unknowns $\theta_{sk}, k = 1, \ldots, K$. The statistics $\langle h_{sl} h_{sk} \rangle$ and $\left\langle \frac{dW_s}{ds} h_{sk} \right\rangle$ can be estimated for all times $t \leq s \leq T$ simultaneously from a single Monte Carlo sampling run using the control u parametrized by θ^0.

The CE method for PI control: Time-independent solution

We consider $\hat{u}(X_s)$ independent of time parametrised by θ. The gradient of the KL divergence involves an integral:

$$\frac{\partial KL(p^*|\hat{p})}{\partial\theta} = \frac{1}{\psi(t,x)}\mathbb{E}_p e^{-S(t,x,u)}\left(\int_t^T ds\,(\hat{u}(X(s)) - u(X(s))) - \int_t^T dW(s)\frac{\partial\hat{u}(X(s))}{\partial\theta}\right)$$

Choosing $u = \hat{u}$ yields the gradient procedure

$$\theta_{n+1} = \theta_n - \eta\frac{\partial KL(p^*|\hat{p})}{\partial\theta_n}\Big|_{u=\hat{u}_n} = \theta_n + \eta\left\langle\int_t^T dW_s\frac{\partial\hat{u}(X(s))}{\partial\theta_n}\right\rangle$$

Example: Linear time-dependent feedback control

For $t_0 \leq t \leq t_1$, the 1-dimensional problem

$$
\begin{aligned}
dX_t =& X_t \left(\frac{dt}{2} + u(tX_t, t)dt + dW_t \right), \\
C =& \mathbb{E} \, \frac{Q}{2} \log(X_T)^2
\end{aligned}
$$

has solution

$$
u^*(t, x) = \frac{-Q \log(x)}{Q(t_1 - t) + 1}.
$$

For the experiments we will take $x_0 = 1/2$, $t_0 = 0$, $t_1 = 1$, $Q = 10$.

Example: Linear time-dependent feedback control

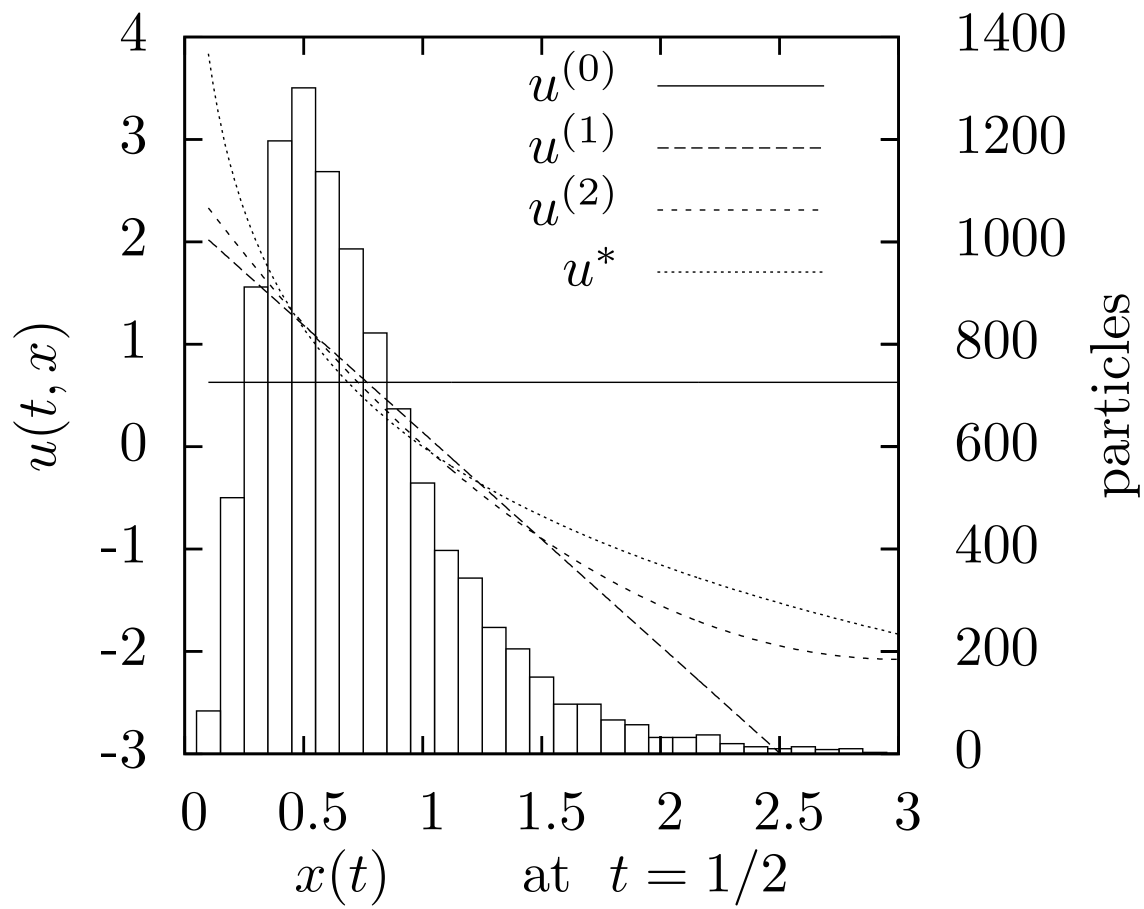
Consider different state-dependent parametrizations:

- one basis function: $\log(x)$ yields exact controller
- three polynomial parameterizations: a constant-, affine- and quadratic-function of the state denoted by $u^{(0)}$, $u^{(1)}$, $u^{(2)}$, e.g. $u^{(2)}(t,x) = a(t) + b(t)x + c(t)x^2$.

	$u=0$	$u^{(0)}$	$u^{(1)}$	$u^{(2)}$	$a(t)\log(x)$	u^*
$\mathbb{E}[S]$	7.526	5.139	1.507	1.461	1.422	1.420
$\mathrm{Var}(\alpha^u)$	1.981	1.376	0.143	0.0506	0.0085	0.0071
FES(%)	34.3	42.08	87.5	95.2	99.1	99.3

Performance estimates of various controllers based on 10000 sample paths.

Example: Linear time-dependent feedback control

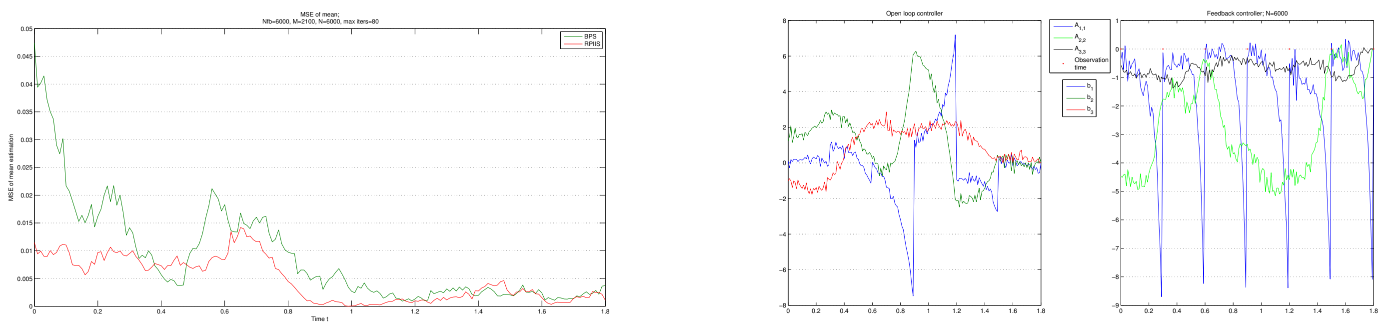


State dependence of the feed-back controllers at the intermediate time $t = 1/2$. The approximate controls were calculated with 10000 sample paths using a time discretization of $dt = 0.001$ for numeric integration. The histogram was created with 10000 draws from $X^{u^*}(t)$ at $t = 1/2$.

Example: Latent state estimation

The path integral control computation is mathematically equivalent to a Bayesian inference problem in a time series model with $p_0(\tau)$ the forward model and $e^{-V(\tau)} = \prod_t p(y_t|x_t)$ is the likelihood of the trajectory $\tau = x_{t:T}|x$. The Bayesian posterior is then given by $p^*(\tau)$.

PICE provides an efficient alternative to particle smoothing methods.



Left: MSE of posterior mean versus time of a chaotic 3-d Lorentz attractor with 7 1-d noisy observations. PI computed $\hat{u}_i(t,x) = \sum_{j=1}^{3} A_{ij}(t)x_j + b_i(t)$ (red) using 80 importance sampling iterations with 6000 particles per iteration. Particle smoothing method (green) using $N = 6000$ forward and $M = 2100$ backward particles. Middle: open loop control b_i versus time. Right: diagonal feedback control terms A_{ii} versus time.

Example: Linear time-independent feedback control

Consider a simple inverted pendulum, that satisfies the dynamics

$$\ddot{\alpha} = -\cos\alpha + u$$

where α is the angle that the pendulum makes with the horizontal, $\alpha = 3\pi/2$ is the initial 'down' position and $\alpha = \pi/2$ is the target 'up' position, $-\cos\alpha$ is the force acting on the pendulum due to gravity. Introducing $x_1 = \alpha, x_2 = \dot{\alpha}$ and adding noise, we write this system as

$$\begin{aligned}
dX_i(s) &= f_i(X(s))ds + g_i(u(s, X(s) + dW(s)) \qquad 0 \leq s \leq T, \quad i = 1,2 \\
f_1(x) &= x_2 \\
f_2(x) &= -\cos x_1 \\
g &= (0,1) \\
C &= \mathbb{E}\int_0^T ds \frac{R}{2} u(s, X(s))^2 + \frac{Q_1}{2}(\sin X_1(s) - 1)^2 + \frac{Q_2}{2} X_2(s)^2
\end{aligned}$$

with $\mathbb{E}dW_s^2 = \nu ds$ and ν the noise variance.

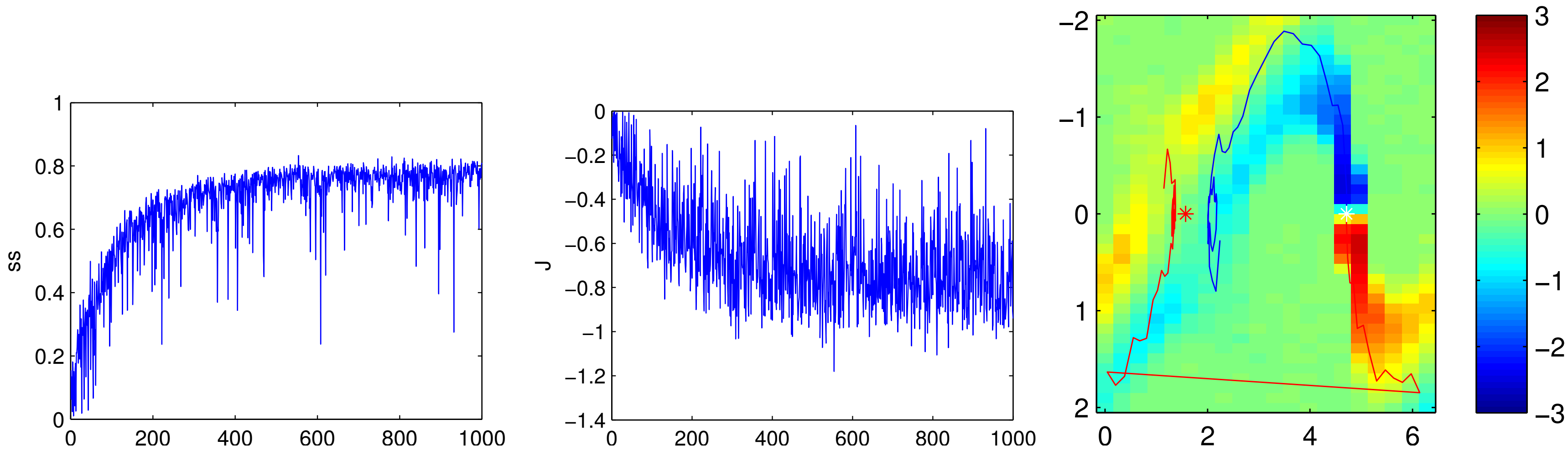
Example: Linear time-independent feedback control

We estimate a time-independent feed-back controller on a grid

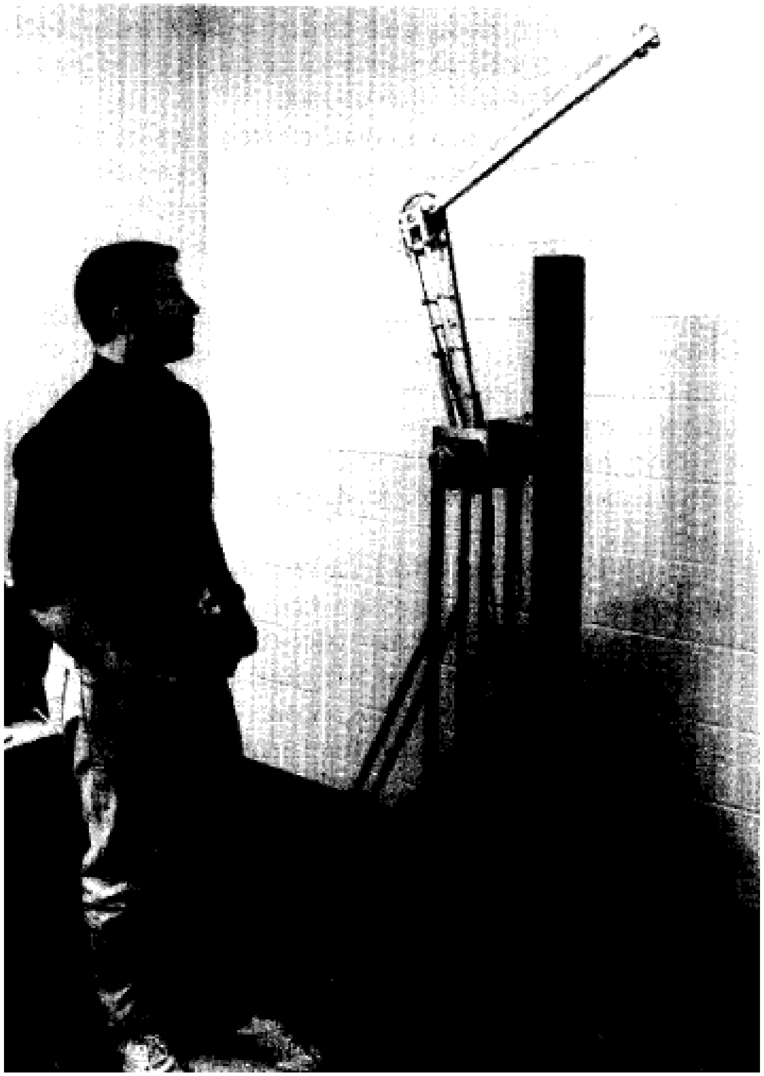
$$\hat{u}(x_1, x_2) = \theta_{k_1,k_2} \ \text{ if } (x_1, x_2) \text{ is in cell } (k_1, k_2)$$

with $k_i, i = 1, 2$ integers that label the grid points.

The results of the path integral learning rule Eq. 1 are shown in fig. **??**.



Acrobot



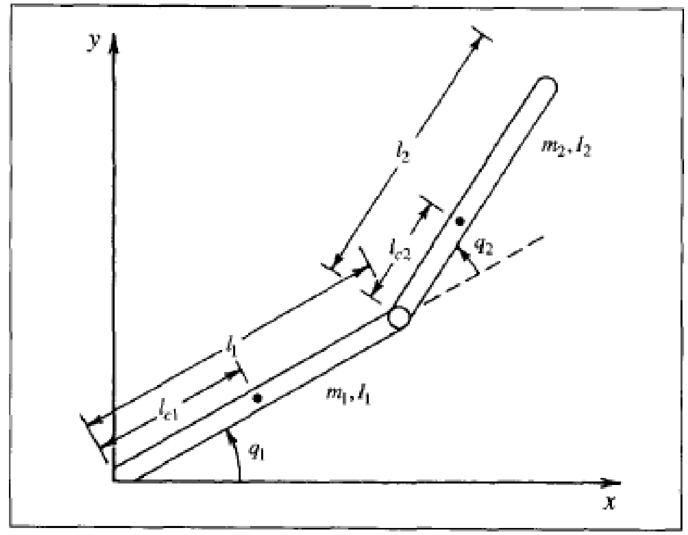


Fig. 1. The Acrobot.

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

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

where

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

Acrobot

(movie92.mp4)

Result after 100 iterations, 50 samples per iteration.

Quadrotors

- circular holding/hovering pattern
  - penalizes large deviations from the centers, collisions and too large/small velocities
  - 15 quadrotor units, rollouts N=7000, horizon H=4
- cat & mouse
  - penalizes large deviations from the mouse, collisions and large/small velocities.
  - Mouse is not controlled and tries to escape the cats

Compute (feed-back) control for current state. Use adaptive importance sampling.

- ≈ 100.000 trajectories/second for 1 second of 1 quadrotor simulation.

UAVs

(AAMAS 2015.mp4)

Kappen et al. 2015

Discussion

PICE presents challenging learning problems, as is evident from the large fluctuations despite the large number of samples for these relatively small problems.

- The weights of the trajectories are proportional to e^{-S} with $S \propto 1/\lambda$ and $\lambda = R\nu$
  - Small λ yields small sample size and difficult learning
  - Large ν requires large controls, requires small R.

  This problem is due to the log transform that is used to linearize the Bellman equation.

- Small deviations from optimallity may yield large decrease in sample size.
  - Optimal model is infinitely large
  - An infinite model requires infinitely many samples to avoid overfitting.
  - for finite samples there is an optimal finite model

Conclusion

Importance sampling improves sampling efficiency:

- optimal control = optimal sampling

Conclusion

Importance sampling improves sampling efficiency:

- optimal control = optimal sampling

Learning state dependent/feedback control with PICE

- CE provides a criterion for parametrized controllers
- learn from self-generated data
- use ∞ data to learn ∞ models
- Connecting Control, Inference and Learning
- application in robotics

Conclusion

Importance sampling improves sampling efficiency:

- optimal control = optimal sampling

Learning state dependent/feedback control with PICE

- CE provides a criterion for parametrized controllers
- learn from self-generated data
- use ∞ data to learn ∞ models
- Connecting Control, Inference and Learning
- application in robotics

Inference:

- reformulate as control problem
- improve estimates through importance sampling controls

S. Thijssen and H. J. Kappen. ”Path Integral Control and State Dependent Feedback.” Phys. Rev. E 91, 032104 Published 2 March 2015

V Gómez, S Thijssen, HJ Kappen, S Hailes ”Real-Time Stochastic Optimal Control for Multi-agent Quadrotor Swarms”. arXiv preprint arXiv:1502.04548, 2015

J Bierkens, HJ Kappen ”Explicit solution of relative entropy weighted control”. Systems & Control Letters 36-43, 2014