An Approximate Inference Approach to Temporal Optimization in Optimal Control

Konrad Rawlik[†], **Marc Toussaint**[⋆], **Sethu Vijayakumar**[†]

[†] *School of Informatics, The University of Edinburgh, UK;* [⋆] *TU Berlin, Germany*

1 – Motivation

- **Stochastic Optimal feedback control** (SOFC) is a plausible movement generation strategy in goal reaching tasks for biological systems → attractive for anthropomorphic manipulators
- For general systems with non-linear dynamics and non-quadratic costs SOFC law can only be found locally and iteratively, e.g., using **iLQG** [1], **AICO** [2] or **DDP** [3]. These algorithms depend on *a priori* specified **temporal parameters** of the task, e.g., time point of reaching a target.
- Choosing **temporal parameters** is non trivial
  - simple distance-duration scaling laws → require knowledge of the movement distance in joint space
  - first exit time formulations → can not account for the influence of subsequent subtasks on previous goals in sequential tasks, e.g. a via-point task.
- Our approach: Express the task in a **canonical time** and jointly optimise the policy and mapping from canonical time to real time → probabilistic formulation leads to an efficient **EM algorithm**.

2 – SOFC Model

Random variables we use here:

$\mathbf{x}_t$	state of a plant (joint angles $\mathbf{q}$ and velocities $\dot{\mathbf{q}}$)
$\mathbf{u}_t$	control signal applied at time t (torques)
r_t	binary task variable indicating success
$\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{r}$	trajectories in $\mathbf{x}, \mathbf{u}$ & r

Probabilities we use here:

plant dynamics	$P(\mathbf{x}_{t+1}\|\mathbf{x}_t, \mathbf{u}_t)$
policy	$\pi(\mathbf{u}_t\|\mathbf{x}_t)$
task likelihood	$P(r_t = 1\|\mathbf{x}_t, \mathbf{u}_t) = \exp\{-cost(\mathbf{x}_t, \mathbf{u}_t, t)\}$
trajectory distribution	$q_\pi(\bar{\mathbf{x}}, \bar{\mathbf{u}}) = P(\mathbf{x}_0)\prod_{t=0}^{T} P(\mathbf{x}_{t+1}\|\mathbf{x}_t, \mathbf{u}_t)\pi(\mathbf{u}_t\|\mathbf{x}_t)$

SOFC inference problem:

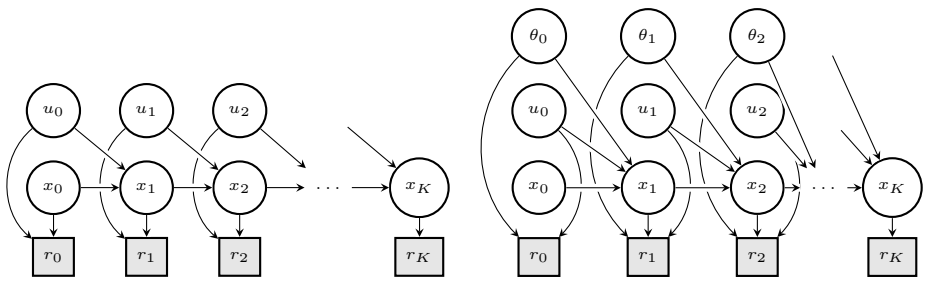
Infer state distribution under a uniform policy π_0, given task success, i.e. $p(\bar{\mathbf{x}}|\bar{r} = 1) = \int_{\bar{\mathbf{u}}} q_{\pi_0}(\bar{\mathbf{x}}, \bar{\mathbf{u}}) \prod_{t=0}^{T} P(r_t = 1|\mathbf{x}_t, \mathbf{u}_t)$

3 – SOFC model with canonical time

Augmented model:

Introduce canonical time τ, s.t. $\tau = \beta(t) = \int_0^t \frac{1}{\theta(s)} ds$ and discretize problem w.r.t. τ.

- $\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{r}$ now indexed in canonical time $1 \dots K$
- New discretized dynamics $P(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{u}_k, \theta_k)$
- New cost $\mathcal{C}(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\theta}) = \sum_{n=0}^{N} \mathcal{J}_n(\mathbf{x}_{\tau_n}) + \sum_{k=0}^{K} \theta_k \mathcal{J}(\mathbf{x}_k, \mathbf{u}_k) + \mathcal{T}(\theta_k)$
  - $\mathcal{V}$ costs incurred at specific time points in canonical time, e.g., reaching a target
  - $\mathcal{J}$ cost incurred throughout the movement, e.g. energy cost
  - $\mathcal{T}$ time mapping cost, e.g., linear cost $\mathcal{T}(\theta) = \alpha\theta$ equivalent to linear duration cost



Augmentation of the standard SOFC model *(left)* leads to the new model *(right)*, note that θ can be seen as an additional control variable.

New SOFC problem:

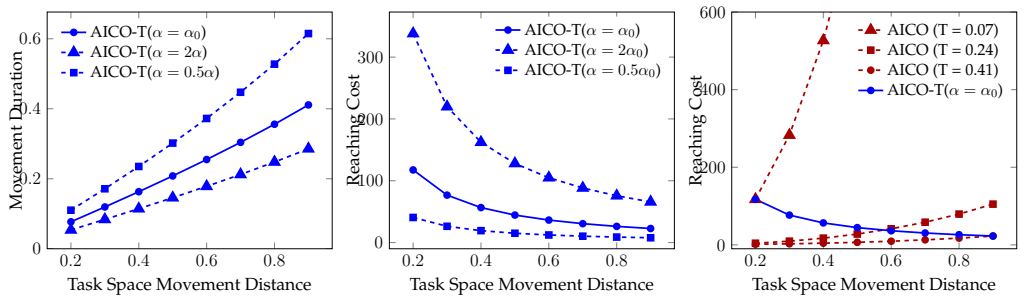
Infer state and time trajectory distribution given success, $P(\bar{\mathbf{x}}, \bar{\theta}|\bar{r} = 1)$.

Algorithm:

Approximate solution, find ML estimate of $\bar{\theta} = \bar{\theta}^*$ and $P(\bar{\mathbf{x}}|\bar{\theta}^*, \bar{r} = 1)$ → use Expectation Maximisation algorithm.

E-Step	find $q_i(\bar{\mathbf{x}}) = P(\bar{\mathbf{x}}|\bar{\theta}^i, \bar{r} = 1)$ → standard SOFC problem, can use iLQG, AICO, DDP, etc.
M-Step	improve $\bar{\theta}^i$ given $q_i(\bar{\mathbf{x}})$ → analytical if q_i Gaussian and $\mathcal{T}$ linear, else gradient step.

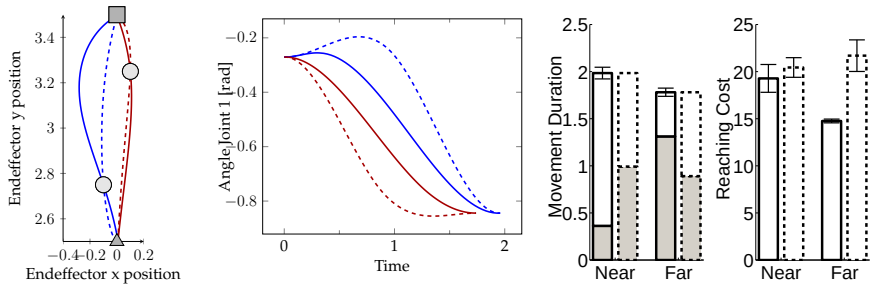
4 – Reaching Task



Standard reaching task with 2dof arm, AICO for approximate E-Step, analytical M-Step.

- We obtain **distance-duration scaling** modulated by time cost parameter α.
- **No task failure** due to inappropriately chosen duration, cf. reaching cost increase with standard AICO.

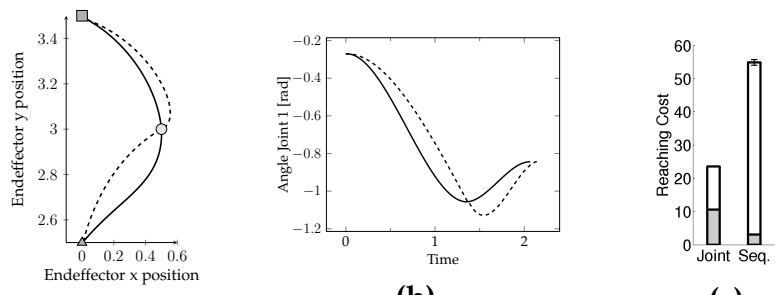
5 – Sequential and via-point tasks



(a) **(b)** **(c)**

Reaching task with intermediate target on 2dof arm, (*Solid*) with temporal optimisation (AICO for approximate E-Step, analytical M-Step) and (*Dashed*) without temporal optimisation (AICO).

- Temporal optimisation leads to non trivial task space trajectories which in joint space are more direct (cf. **(a)** & **(b)**, n.b. joint trajectory of only one joint shown).
- Overall movement **duration split** according to target distances in joint space (cf. **(c)** left).
- Temporal optimisation leads to **lower costs** (cf. **(c)** right).



(a) **(b)** **(c)**

Optimisation over entire task (*Solid/Joint*) compared to sequentially optimising for subtasks, i.e., sequential first exit time, (*Dashed/Seq*).

- Joint optimisation leads to **smoother trajectories** as subsequent tasks influence behaviour in the current task (cf. **(a)** & **(b)**, n.b. joint trajectory of only one joint shown).
- Joint optimisation leads to significantly **lower costs** (cf. **(c)**)

References

[1] Li W. and Emanuel Todorov, E. *A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems.* Proceedings of the American Control Conference, 2005

[2] Toussaint, M. *Robot Trajectory Optimization using Approximate Inference.* ICML, 2009

[3] Jacobson, D. and Mayne, D. *Differential Dynamic Programming.* Elsevier, 1970

Acknowledegments:

Contact Author: Konrad Rawlik (k.c.rawlik@ed.ac.uk) for further details on this work.