On Optimal Control and Expectation-Maximisation: Theory and an Outlook Towards Algorithms

Tom Lefebvre
Dynamic Design Lab,
Department of Electromechanical, Systems and Metal Engineering
Faculty of Engineering and Architecture, Ghent University
Technologiepark 131, 9052 Zwijnaarde, Belgium
E-mail: tom.lefebvre@ugent.be

June 2022

Abstract. In this work we demonstrate how both the Stochastic and Risk Sensitive Optimal Control problem can be treated by means of the Expectation-Maximisation algorithm. We show how such a treatment materialises into two separate iterative programs that each generate a unique but closely related sequence of density functions. We motivate to interpret these density functions as beliefs, ergo as probabilistic proxies for the deterministic optimal policy. More formally two fixed point iteration schemes are derived with the stationary point coinciding with the deterministic optimal policies on behalf of the proven convergence of Expectation-Maximisation methods. We are inclined to point out our results are intimately related with the paradigm of Control as Inference. Control as inference here refers to a collection of approaches which aim is also to recast optimal control as an instance of probabilistic inference. Although said paradigm already resulted in the development of several powerful Reinforcement Learning algorithms, the fundamental problem statement usually is introduced by teleological arguments. We argue that the present results demonstrate that earlier established Control as Inference frameworks in fact isolate a single step from either of the proposed iterative programs. In any case the present treatment provides them with a deontological argument of validity. By exposing the underlying technical mechanism we aim to contribute to the general acceptance of Control as Inference as a framework superseding the present Optimal Control paradigm. In order to motivate the general relevance of the presented treatment we further discuss parallels with Path Integral Control and other areas of research before sketching the outlines of future algorithmic development.

1 Introduction

The issues posed by a desire to control dynamic systems is a research activity that concerns many engineering disciplines. In particular for systems that are high dimensional, nonlinear and continuous in the system state-space, finding a satisfactory control remains one of the most challenging problems faced by their respective research communities. Formally it is aimed to find a control policy, which maps system outputs to system inputs, so that the closed-loop system is both stable, and, often, exhibits rich and complex behaviour or dynamic skills such as locomotion or manipulation.
A well known paradigm to synthesise such policies is provided by Optimal Control. Here the policy is determined such that when it is used to control the system, the ensemble is expected to accumulate a minimised cost over a finite (or infinite) time horizon. By proper design of the corresponding cost functions it is possible to encode abstract behavioural and dynamic features into the policy. Finding such optimal policies is an elegant and appealing theoretical concept but is met by significant difficulties in practice and explicit expressions for the optimal policy exist only for a handful of problem statements.

In an effort to address these issues, some researchers have turned to probabilistic methods, attempting to rephrase the problem of deterministic optimisation as one of probabilistic inference. It is possible to view these endeavours as part of a larger movement referred to as probabilistic numerics [4, 5]. Algorithms for numerical problems, such as optimisation, proceed iteratively and typically maintain a deterministic running estimate of the correct solution. With every iteration that estimate is improved based on any information that comes at the result of probing the functions that govern the problem. Take for example gradient based optimisation methods where local function and gradient evaluations are processed into a step update of the point estimate. In contrast, probabilistic numerics pursues methods that, in place of such point estimates, update probability measures over the space of possible solutions. Note that the manifestation and interpretation of probability is here in the first place epistemic and representative for missing information in a computation that itself is otherwise entirely deterministic.

When pursued in the context of (optimal) control, these attempts are usually referred to with the umbrella term Control as Inference (CaI) [6, 7, 8]. In part, they consolidate dualities and connections between control and estimation that have captivated researchers for decades and which have resurfaced in recent theoretical work, the majority of which was contributed by the Reinforcement Learning (RL) community. From a practical point of view, they give rise to new theoretical paradigms that hopefully will bring forth original algorithms that can draw from the mature (approximate) inference machinery developed by the Machine Learning community.

1.1 Contributions and paper organisation

Essentially we explore the relation between the governing principles associated to Control as Inference and the actual optimal control problems that lie underneath.

- In section 3 we reformulate the Stochastic and Risk Sensitive Optimal Control problem so that they can be treated by means of the Expectation-Maximisation algorithm. This treatment suggests two convergent iterative programs which generate a sequence of policy belief functions.
- In section 4 we highlight particular features and discuss parallels with Path Integral Control and other areas of research. We further specialise the programs to the Linear- and the Linear-Exponential-Quadratic-Regulator.
- In section 6 we sketch outlines for future algorithmic development.

2 Preliminaries

We review here a number of topics that are required to develop the main results in section 3. The treatment of optimal control is restricted to discrete time. The results further apply to both discrete and continuous state and action spaces.
2.1 Optimal control problems

Let us begin with defining the two optimal control problems that, as we will show, can be treated by means of the Expectation-Maximisation algorithm. Before we venture down that road, first we introduce a probabilistic model that will be of great assistance in stating the problems.

The mathematical stage where the subjects of this work materialise is provided by Controlled Markov Chains (CMCs). A CMC can be depicted using the following probabilistic graph model. Here $t$ represents discrete time, $T$ is some natural number associated to the total length of sequences. The state variables, $x_0, ..., x_T$, assume values in the set $X$, the control variables or actions, $u_0, ..., u_T$, assume values in the set $U$. The sets $X$ or $U$ can be either discrete or continuous. For notational convenience we further define the tuple $\xi = (x, u)$ and the notation formats $x_t = \{x_0, ..., x_t\}$ and $π_t = \{x_t, ..., x_T\}$ which we use to refer concisely to a leading or trailing part of a sequence with $t$ referring to the final or initial time instance of the corresponding subsequence. We silently assume that a complete sequence starts at time $t = 0$ and ends at time $t = T$.

![Graphical representation of a Controlled Markov Chain.](image)

CMCs are closely related to Markov Decision Processes. The principal difference with a Markov Decision Process is that, for the time being, the choice of action, $u_t$, as a function of the present state, $x_t$, is not governed by the intention of rendering some utility function optimal. Instead, we assume that this choice itself is subject to an uncertain decision function or policy.

To specify the global probabilistic model, we need simply to characterise the local uncertainty models. The model, $M$, is as such defined by the initial model, $p(x_0)$, a transition model, $p(x_{t+1}|x_t, u_t)$ and a policy model, $π_t(x_t, u_t) = p(u_t|x_t)$. For the remainder we will focus on continuous spaces so that all models can be characterised using a probability density function, though the translation to discrete spaces and probability distributions is straightforward. Throughout we refer to the set of all feasible probability density functions with $P$. We will rely on the element that is preceding the relation symbol, ‘$∈$’, to imply the arguments of the corresponding sets. Taker for example, $π_t ∈ P$, which implies that we refer to the set of functions, $P$, with argument $x_t$ that normalise over the argument $u_t$. The fact that these sets may differ depending on $x_t$, is relegated as an obsolete verbosity to keep a check on the notational overload. As specified, the policy models therefore determine the uncertainty about the action, $u_t$, given that the corresponding state assumes a value, $x_t$.

By imposing the standard Markov conditions on the probabilistic graph in Fig. the joint model is completely determined by its local models. Specifically, the uncertainty about a trajectory, $ξ_T$, is determined by the product of all local transition and policy models.

\[\xi_T = π_T(x_T, u_T) \cdot π_{T-1}(x_{T-1}, u_{T-1}) \cdot ... \cdot π_0(x_0) ≡ π_T(x_T)\]

We note that whenever a cycle is present in such a graph model, the associated variables could also be concatenated. This simplifies the model at the cost of having no direct access to the probability that governs their relation except after marginalisation.
On Optimal Control and Expectation-Maximisation

\[ p(\xi_T, \pi_T) = p(x_0) \prod_{t=0}^{T-1} p(x_{t+1} | x_t, u_t) \pi_t(\xi_t) \]  \hspace{1cm} (1)

Given the crucial role that the policy models will play throughout; we say, and emphasise notationally, that the trajectory density is parameterised by the policy sequence \( \pi_T \).

2.1.1 Stochastic optimal control  The first problem that we consider in this work is that of Stochastic Optimal Control (SOC) (or Markov Decision Processes for discrete state-action spaces). The goal is to render some cost or utility function, \( R \), extreme by administering a sequence of carefully selected policies, \( \pi_T \in \mathcal{P} \). This problem formalises the setting for many methodologies tailored to control, amongst which are trajectory generation, model based predictive control, reinforcement learning, etc.

We can use the parameterised CMC trajectory model to formulate the problem concisely.

\[ \pi_T^* = \arg \min_{\pi_T \in \mathcal{P}} \int p(\xi_T, \pi_T) R(\xi_T) d\xi_T \]  \hspace{1cm} (2)

The cost function \( R \) denotes an accumulated cost or cost-to-go. The functions, \( r_t \) and \( r_T \), represents the cost rate and final cost and can be used to encode abstract behavioural and dynamical features into the control problem.

\[ R(\xi_T) = r_T(x_T) + \sum_{t=0}^{T-1} r_t(\xi_t) \]  \hspace{1cm} (3)

In what follows, we will occasionally rely on the argument of \( R \) to imply the terms in the sum above. For example, \( R(\xi_t) \), would refer to the leading terms, whilst \( R(\xi_T) \) would refer to the trailing terms in (3).

2.1.2 Risk sensitive optimal control  The second problem that enjoys our interest is that of Risk Sensitive Optimal Control (RSOC). Following the observation that for affine system models (e.g. see section 4.5) the solution to the SOC problem and its deterministic counterpart are equivalent, a risk sensitive version has been proposed where instead of minimising the cumulative performance criteria \( R \), the exponential of \( R \), \( \exp(-\sigma R) \), is maximised [9].

Thinking of the trajectory density as a collection of alternative histories, using such an exponential utility function puts less (or more) emphasis on the successful histories. We refer to the work of Whittle [10, 11, 12] for an excellent exhibition on RSOC (and the use of utility functions in decision making in general).

\[ \pi_T^\sigma = \arg \min_{\pi_T \in \mathcal{P}} \frac{1}{\sigma} \log \int p(\xi_T, \pi_T) e^{-\sigma R(\xi_T)} d\xi_T \]  \hspace{1cm} (4)

The solution is referred to as risk seeking or averse for \( \sigma < 0 \) either \( \sigma > 0 \). In this work we focus on the risk averse version but note that the risk seeking version can be treated in the same way as will be presented here. Further, it can be shown that for the limit \( \sigma \to 0 \), the RSOC collapses onto the SOC problem. This result is characteristic of the fact that the solution of the SOC problem is invariant to scaling of the cost function \( R \) whilst the solution of the RSOC problem is not. For conciseness we further absorb \( \sigma \) into \( R \) which can be achieved by appropriate scaling of the functions \( \pi_T \).
2.1.3 Some additional remarks

In the previous two paragraphs we have formulated two regular optimal control problems making use of the trajectory density \( p(\xi_T; \pi_T) \) associated to CMCs. Recall that therefore the trajectory density was parameterised by some arbitrary sequence of policy models, \( \pi_T \), living in the set of densities conditioned on \( \xi_T \). This is not the standard approach in the control literature and may appear to suggest that the solution of either of these problems is governed by an uncertain sequence of decision functions. We strongly argue against this suggestion and emphasise that in fact neither problem is rendered extreme by an uncertain policy. The solution is given by a sequence of optimal deterministic policy functions, \( \pi^*_t : X \mapsto U \). It is intuitively clear that the objective function cannot be rendered extreme by gambling on the outcome of a control \( u_t \) given a state \( x_t \). Regardless both problems can be treated probabilistically, in the sense that we substitute a policy density, with the optimal solution collapsing onto a Dirac delta function. More formally we have that the set of all deterministic decision functions is a subset of the set of all uncertain decision function. As a result, we only extend the optimisation space. This representation will prove useful to our purpose later on.

Second, both problems provide us with the same solution when the underlying dynamics are deterministic. One will agree that the way of thinking of the trajectory distribution as a collection of alternative histories then collapses onto a single history, rendering the influence of the exponentiation obsolete. In that same line of thought it follows that the solution of a deterministic optimal control problem is uniquely defined by the optimal control sequence \( u^*_T \) and associated optimal state sequence \( \xi^*_T \) for given and fixed initial state, \( x_0 \). As such the optimal trajectory is defined as an instantiation of the underlying optimal policy given the initial state \( x_0 \) so that \( u^*_t = \pi^*_t(x^*_t) \). For stochastic systems, the solution cannot be a unique trajectory given that the state transitions differently for every history and we must adapt the course of actions accordingly.

In conclusion we include a summary of common numerical solution strategies. Therefore, we must first note that both problems exhibit optimal substructure and can be recast using the principle of dynamic programming into \( T \) smaller optimisation subproblems that are backward recursive dependent. As such it is possible to construct backward recursive equations that govern the optimal policy. For conciseness, we have not included these governing equations here. They are included in Appendix A. Although the recursive statement of the problem yields beneficial insight in the problems nature, these backward recursive sequence of optimisation subproblems still cannot be solved for general models. Only a small subset of problem formulations can be treated in this way. For continuous spaces, this problem can only be solved analytically in case of the Linear-Quadratic-Regulator (LQR). Again, we refer to Appendix A. To treat nonlinear optimal control problems a wide variety of iterative numerical procedures have been developed better known as trajectory optimisation algorithms. Amongst them are the iterative LQR (iLQR) \cite{13}, Differential Dynamic Programming (DDP) \cite{14, 15, 16, 17}, Direct Multiple Shooting (DMS) \cite{18, 19} and hybrids between the DDP and DMS \cite{20, 21, 22}. At the heart of all of these algorithms lies the LQR solution. The dynamics are linearised (with the exception of the DDP algorithm where also second order derivative of the dynamics are used) and a quadratic approximation of the cost is constructed about some reference trajectory \( \xi^*_T \). The iterative LQR policy is then calculated about the reference \( \xi^*_T \), a closed-loop simulation of the dynamics is performed to find the next iterate trajectory \( \xi^{t+1}_T \). And so on.
These approaches are generally in line with conventional gradient-based solution methods for static optimisation problems where a point estimate of the solution is maintained over the iterations of the algorithm.

2.2 Probabilistic inference

In this section we will review some basic results from probabilistic inference. They are similar in that they allow to make precise revisions of our uncertainty in light of new evidence [24].

For that purpose, in this paragraph we briefly consider an arbitrary probabilistic model, $M$, that describes the uncertain relation between all the variables that make up the model, similar to the CMC. An experiment refers to the situation where all these variables assume a deterministic value according to the local probabilities that govern the global model. For convenience we cluster all variables of the model into two sets. The set, $X \in X$, denotes all variables that are hidden and cannot be evaluated during an experiment. The set, $Z \in Z$, denotes all observable variables, meaning they assume a measurable value during an experiment. Such a general model is governed by two local models. The local model, $p(X)$, often referred to as the prior model, and, the emission model, $p(Z|X)$, which expresses the probability of observing a specific, $Z$, for different values of $X$.

2.2.1 Bayesian Inference

Broadly speaking Bayesian Inference (BI) refers to the practice of computing conditional probabilities in arbitrary probabilistic models according to the rules of probability. Bayes’ rule is used to convert the prior probability, $p(X')$, capturing our initial uncertainty – that is therefore based solely on the probabilistic model, $M$ – about a subset of unobservable random variables, $X' \in X' \subset X$, into a posterior probability, $p(X|Z)$, by incorporating the evidence provided by the set of observable random variables. The probabilities, $p(X')$, and, $p(X'|Z)$, are said to reflect our uncertainty about the hidden subset, $X'$, prior and posterior to an experiment.

$$p(X'|Z) = \frac{p(Z|X')p(X')}{p(Z)} \propto p(Z|X')p(X') \quad (5)$$

A simplifying and therefore key feature is the concept of conditional independence which often makes it possible to make inferences in arbitrarily complex probabilistic models by decomposing the local models, $p(X)$, and, $p(Z|X)$, based on the causal independence relations that exist between the elements of $X$ and $Z$. Fundamental applications in estimation are filtering and smoothing. Some useful related results are given in [Appendix B]

2.2.2 Entropic Inference

BI allows to process evidence that is represented by the outcome of experiments. We can arguably refer to information sources of that kind as empirical evidence. Entropic Inference (EI) [25, 26] is an extended probabilistic concept with deep roots in information-theory, however described as a fundamental aspect of probability theory by Jaynes and others [23, 27]. EI allows to treat

§ Here we embrace Jaynes’ view of probability theory as extended logic and his interpretation of probability distributions and densities as carriers of information [23]. In that view, predictions made by exerting the probabilistic model are subjective to the entity that is calling upon them, and more specifically the information that is available to that entity.
information represented by constraints that affect our belief space. We refer to these information sources as structural evidence. Here an interesting situation is met. Since the evidence does not directly relate to any variables of the model, a different line of approach is needed. We can formalise the situation as follows. Evidence has been presented that assures that the posterior, \( \pi \), is constrained to some subset \( P^* \subseteq P \).

The principle that facilitates EI is that of minimum relative entropy \([28, 29]\) which in turn are extensions of Laplace’s or Jaynes’s principle of insufficient reason \([30]\). In words it states that the unique posterior belief, \( \pi \), living in the constrained density set, \( P^* \), is the one that is hardest to discriminate from the prior, \( \rho \). The measure that is used to discriminate between two beliefs is the relative entropy.

\[
\mathbb{D}[\pi || \rho] = \int \pi(X) \log \frac{\pi(X)}{\rho(X)} dX
\]

The posterior \( \pi \) can then be determined by minimising the relative entropy whilst constraining the solution to the set \( P^* \). Jaynes’ principle then asserts that \( \pi \) is the least-biased posterior that does satisfy the constraints.

\[
\pi^* = \underset{\pi \in P^* \subseteq P}{\arg \min} \mathbb{D}[\pi || \rho]
\]

The relative entropy provides a measure of the inefficiency of assuming that the uncertainty is modelled by \( \rho \) when it should be \( \pi \). The relative entropy is asymmetric in its arguments, always positive and zero if and only if \( \pi \equiv \rho \). This last property implies that if \( \rho \in P^* \), there is in fact no reason to update the prior and the information is redundant. When the prior is a uniform, the concept collapses onto the Maximum Entropy (ME) principle \([31]\).

A common constraint is the following

\[
P^f = \left\{ \pi \in P \left| \int \pi(X) f(X) dX = 0 \right. \right\}
\]

The corresponding posterior is known as the MaxEnt density \([23]\). Note that this posterior is also proportional to the prior multiplied with a likelihood which is, remarkably, analogous to the Bayesian update. Here \( \lambda \) is a Lagrangian multiplier that asserts the constraint is satisfied. Note that the distribution needs to be normalised and that normalisation constant, \( \eta \), also depends on \( \lambda \).

\[
\pi^f(X) \propto e^{-\lambda f(X)} \rho(X)
\]

In information-theory the operation in (7) is sometimes referred to as the Information or I-projection of the prior \( \rho \in P \) onto the constrained distribution space \( P^* \subseteq P \) \([27, 32, 33]\). Given that the relative entropy is asymmetric in its arguments, a reciprocal projection exists which is known as the Moment or M-projection of the prior \( \rho \) onto the distribution space \( P^* \subseteq P \) \([32, 33]\). In particular

\[
\pi^\ast = \underset{\pi \in P^* \subseteq P}{\arg \min} \mathbb{D}[\rho || \pi]
\]

The I-projection typically under-estimates the support of \( \rho \) and locks onto its principle modes because \( \pi \) should equal zero when \( \rho \) does, to ensure the relative entropy stays finite. The M-projection on the other hand typically overestimates the support of \( \rho \). That is because \( \pi(X) > 0 \) when \( \rho(X) > 0 \).

A more elaborate way of viewing this is that the relative entropy quantifies the average additional amount of information required to decode the message \( X \) and using some optimal decoding scheme, assuming \( X \) is distributed according to \( \rho \) whilst instead \( X \) is distributed according to \( \pi \) \([27]\).
2.2.3 Expectation-Maximisation

In this third and final paragraph we cover the main ideas from the Expectation-Maximisation (EM) algorithm. Essentially this algorithm reduces to an iterative solution technique for finding Maximum Likelihood Estimates (MLE) of probabilistic models with latent variables [24, 32]. As opposed to the inference techniques discussed so far, it is used to find a point estimate of probabilistic model parameters not so much a posterior belief. The main trick relies on a useful decomposition that recollects elements from either other inference mechanism.

Reconsider the probabilistic model with hidden and observable variables, $X$ and $Z$. Further suppose that the model, $M$, is characterised by a set of variables $\theta$, so that the joint density is given by, $p(Z,X;\theta)$. The goal is thus to determine a probabilistic model, $M_\theta$, through identification of the parameters, $\theta$.

The MLE approach is to select the model, $M_\theta$, that maximises the likelihood of observations, $Z$

$$\max_\theta J(\theta)$$

where

$$J(\theta) = \log p(Z;\theta) = \log \int p(Z|X;\theta)p(X;\theta)dX$$

It is well-known that this objective is difficult to treat on account of the integral expression, provided that $X$ is unobserved and the distribution $p(X)$ is unknown before attaining a value for the parameter $\theta$.

This brings us to the main trick of EM. To circumvent the intractable inference, an auxiliary inference density, $q(X)$, is introduced. The inference distribution allows to decompose $J$ into another surrogate objective, $L$, – often referred to as the evidence lower bound (ELBO) –, and a relative entropy error term. Note that since the relative entropy term is always positive, the ELBO, $L$, bounds $J$ from below.

$$\log p(Z;\theta) = \int q(X) \log \frac{p(Z,X;\theta)}{q(X)}dX + \int q(X) \log \frac{q(X)}{p(X|Z;\theta)}dX$$

This problem decomposition allows to tackle the original optimisation problem in two steps which can be iterated. This iterated two step procedure is referred to as the EM algorithm. (E-step) First the error term is minimised to find the optimal inference density, $q$, fixing parameters, $\theta$. (M-step) Second the ELBO, $\mathcal{L}$, is maximised to find the optimal parameter values, $\theta$, fixing the variational inference distribution, $q$, and so on. Suppose for example that the current value of the parameters is $\theta^{old}$. It is easily verified that the E-step solves for the posterior density, $p(X|Z;\theta^{old})$. In the M-step the inference density, $q$, is fixed and the ELBO is maximised rendering some new value, $\theta^{new}$. By definition, selecting $\theta$ to maximise $\mathcal{L}$, improves $J$ at least as much as $\mathcal{L}$ improved. By repeatedly performing the E- and M-step, we eventually reach a stationary value of $\theta$, guaranteed to be a local maximum of $J$ for sufficiently well-behaved problems.

**E-step** Minimise the error term w.r.t. $q$ fixing $\theta$.

**M-step** Maximise the ELBO w.r.t. $\theta$ fixing $q$.

Finally note that the E-step can also be understood as maximising the ELBO for fixed $\theta$ provided the decomposition below. As such the EM algorithm can also be viewed as an example of a coordinate ascent algorithm on the ELBO.

$$\mathcal{L}[q(X)|Z;\theta] = -\mathcal{D}[q(X)||p(X|Z;\theta)] + \log p(Z;\theta)$$
3 Expectation-Maximisation of Optimal Control problems

The goal of this section is simply to demonstrate that either of the optimal control problems introduced in section 2.1 can be addressed by means of the EM algorithm. We detail our approach in two separate subsections, addressing one problem at a time.

As it turns out either problem reduces to a sequence of what are essentially information-theoretic reformulations of the root problem, using the previous solution as prior for the new solution. Formally therewith we establish a fixed point iteration with the stationary point coinciding with the deterministic optimal control. The benefit of using the EM approach to arrive at these programs is that convergence of the sequence is an inherited property.

3.1 Regularised Stochastic Optimal Control

Reconsider the SOC problem defined in (2). We start our analysis by decomposing the problem as given below. Although lacking a straightforward probabilistic motivation, this decomposition matches the structure required for application of the EM algorithm.

\[
\int p(\xi_T; \pi_T)R(\xi_T)d\xi_T = \int p(\xi_T; \pi_T) \log \frac{1}{e^{-R(\xi_T)}}d\xi_T = \int p(\xi_T; \pi_T) \log \frac{p(\xi_T; \pi_T)}{p(\xi_T; \rho_T)e^{-R(\xi_T)}}d\xi_T
\]

(15)

The optimal control framework imposes the structure of the controlled trajectory density, \( p(\xi_T; \pi_T) \), onto the inference density, \( q \). Consequently, we do not need to adopt the less informed general form prescribed by section 2.2.3 and may view \( \pi_T \) as parametrising the density \( q(\xi_T) \). Further we may view \( \rho_T \) as parametrising the ELBO. Application of the EM-algorithm on the decomposed problem thus entails iterating the following two steps.

**E-step** Minimise the error term for \( \pi_T \) keeping \( \rho_T \) fixed.

**M-step** Maximise the ELBO for \( \rho_T \) keeping \( \pi_T \) fixed.

Solution of the M-step is here trivial. The distribution that renders the ELBO minimal is \( \rho_T \equiv \pi_T \). Although the solution here is straightforward, the use of the EM algorithm does induce an iterative mechanism to treat the underlying optimal control problem.

The question here reduces to solving the E-step.

\[
\pi_T^* = \arg \min_{\pi_T \in \mathcal{P}} \mathbb{E}_{\xi_T \in \mathcal{P}} \left[ \| p(\xi_T; \pi_T) \| p(\xi_T; \rho_T)e^{-R(\xi_T)} \right]
\]

(16)

3.1.1 Explicit solution of the E-step

First verify that problem (16) can be recast as

\[
\min_{\pi_T \in \mathcal{P}} \mathbb{E}_{\xi_T \in \mathcal{P}} p(\xi_T; \pi_T) \left[ R(\xi_T) + \log \frac{\pi_T(\xi_T)}{\rho_T(\xi_T)} \right]
\]

(17)
We can treat this problem as a variational optimisation problem. Note that the problem exhibits an optimal substructure, which can be exploited to solve it more elegantly.

Consider

$$\min_{\pi_t \in \mathcal{P}} \int p(\xi_T; \pi_T) \left( R(\xi_T) + \frac{\pi_T(\xi_T)}{\rho_T(\xi_T)} \right) d\xi_T$$

$$= \min_{\pi_t \in \mathcal{P}} \int p(\xi_t, x_{t+1}; \pi_T) \left( R(\xi_t) + \log \frac{\pi_T(\xi_t)}{\rho_T(\xi_t)} \right) V^*_t(x_{t+1}) dx_{t+1}$$

where

$$V^*_t(x_t) = \min_{\pi_t \in \mathcal{P}} \mathbb{E}_{p(\xi_t|x_t; \pi_t)} \left[ R(\xi_t) + \log \frac{\pi_t(\xi_t)}{\rho_t(\xi_t)} + \mathbb{E}_{p(x_{t+1}|\xi_t)} [V^*_{t+1}(x_{t+1})] \right]$$

Further note that function, $V^*_t$, satisfies a recursion, where $V^*_T(x_T)$ is initialised as $\exp(-r_T(x_T))$

$$V^*_t(x_t) = \min_{\pi_t \in \mathcal{P}} \mathbb{E}_{p(\xi_t|x_t; \pi_t)} \left[ r_t(\xi_t) + \log \frac{\pi_t(\xi_t)}{\rho_t(\xi_t)} + \mathbb{E}_{p(x_{t+1}|\xi_t)} [V^*_t(x_{t+1})] \right]$$

The variational optimisation problem governing, $V^*_t$, can be solved explicitly by making use of variational calculus. Therefore, the complete solution of (16) can be cast into a backward recursion, governed by the following three equations.

$$\pi^*_t(x_t | u_t) = \rho_t(\xi_t) \frac{\exp(-Q^*_t(\xi_t))}{\exp(-V^*_t(x_t))}$$

$$V^*(x_t) = -\log \mathbb{E}_{p(u_t|x_t; \rho_t)} [\exp(-Q^*_t(\xi_t))]$$

$$Q^*(x_t) = r_t(\xi_t) + \mathbb{E}_{p(x_{t+1}|\xi_t)} [V^*_t(x_{t+1})]$$

It is easily verified that these expressions allow to determine a sequence of policy densities, $\{\pi^*_t\}_{t \in \mathbb{N}}$. By practising the substitution, $\rho_T \leftarrow \pi_T^*$, and repeated application of the equations in (21), a program is specified to solve the variational optimisation problem in (16) explicitly.

### 3.1.2 Alternative motivation of (16)

We can arrive at the same result from the previous paragraph using a different line of inquiry. It serves our development right to include such an alternative argument as it provides a principled probabilistic motivation for problem (16) where, in a way, we lacked one earlier on. Furthermore, it attributes new significance to the principle of EI as a technical instrument that extends its usual purpose of assigning prior densities based on structural evidence.

Consider the problem of minimising the function, $f : \mathcal{X} \rightarrow \mathbb{R}$. We assume, $f$, to possess a single global minimum, $x^* = \arg \min_{x \in \mathcal{X}} f(x)$, for ease of exposition. We initiate our argument with the observation that classic numerical optimisation methods iterate a point estimate, say $x^g$, of the solution $x^*$. With every iteration, additional information, $I$, is retrieved based on which the point estimate, $x^g$, is updated to a new point estimate, say $x^{g+1}$. Now instead of focusing on a point estimate of the isolated solution, we could also model this iterative search procedure with a sequence of beliefs $\{\pi^g\}_{g \in \mathbb{N}}$ (over iterations, not time) with every $\pi^g$ expressing our uncertainty about any of the possible solutions. This agrees with the main ideas...
advocated by probabilistic numerics. The more information that is gathered over the iterations, the more certain we are about the solution. If we can establish a sequence that improves monotonically on the objective so that \( E_{\pi^{g+1}}[f] < E_{\pi^g}[f] \), by definition the beliefs converge to a Dirac delta distribution centred at \( x^* \).

To establish such a sequence, we require a consistent inference procedure that facilitates an update of the form \( \pi^{g+1} \leftarrow \pi^g \). To construct such an inference procedure, we can make use of the principle of minimal relative entropy or thus EI. Since a priori no or little information about the possible solutions is available, this situation can be represented mathematically by encoding uncertainty about the solution in a prior probability density, \( \pi^0 \), which support covers the feasible solution space, \( X \), and assigns equal probability to each solution. Then to gradually decrease the uncertainty over iterations, a posterior belief, \( \pi^{g+1} \), is desired that discriminates least from the prior belief, \( \pi^g \), but enough so that the expected objective function value produces a lower estimate on the expected cost. Then following the interpretation of the relative entropy as a measure of the inefficiency of assuming that the density is \( \pi^g \) when the true density is \( \pi^{g+1} \) [27], we could use the I-projection to project \( \pi^g \) onto the belief space defined by the constraint, \( E_{\pi}[f] + \Delta \leq E_{\pi^g}[f] \). In other words, a belief space containing distributions that perform better on \( f \) than the prior. By definition this, update mechanism produces a monotonically decreasing sequence. The problem that we solve is thus defined as \( \min_{\pi} D[\pi \parallel \pi^g] \) s.t. \( E_{\pi}[f] + \Delta \leq E_{\pi^g}[f] \).

First we might want to note that the sequence \( \mathcal{P}^{g+1} = \{ \pi \in \mathcal{P} \mid E_{\pi}[f] + \Delta \leq E_{\pi^g}[f] \} \) is logically consistent, with \( \mathcal{P}^{g+1} \subset \mathcal{P}^g \) as long as \( \Delta \) is chosen carefully. As such, no iteration is redundant nor will any iteration generate the empty set. Second, the issue needs to be addressed, that in practice we cannot desire to keep on subtracting some arbitrary amount \( \Delta \) from the expected cost every iteration. This issue is resolved by considering the explicit solution of this update procedure. By using the method of Lagrangian multipliers it is easily verified that \( \pi^{g+1} \propto \pi^g \cdot e^{-\lambda f} \) where \( \lambda \) must be determined so that the constraint is satisfied. It is easily shown that \( l(\lambda) = E_{\pi^{g+1}}[f] \) is monotonically decreasing as a function of \( \lambda \) by verifying that the gradient, \( l'(\lambda) = -\text{cov}_{\pi^{g+1}}[f] > 0 \), is negative for \( \lambda \geq 0 \). Hence by choosing \( \lambda > 0 \) instead of \( \Delta \) we are certain to have attained the same goal [34] [35]. The larger we choose the value of the multiplier, the faster the implied sequence will converge to the underlying deterministic solution. It also follows that when the sequence converges, that is, when it reaches the minimal expected cost attainable, it must coincide with the Dirac delta centred at the true solution \( x^* \). This follows directly from the definition of the minimum and the natural shrinking of the sets \( \mathcal{P}^g \).

The same argument can be advanced for problem (2). Now we need to solve the following variational optimisation problem:

\[
\min_{\pi_T} D[p(\xi_T; \pi_T) \parallel p(\xi_T; \pi^g_T)] \quad (22a)
\]

subject to

\[
\int p(\xi_T; \pi^g_T) R(\xi_T) d\xi_T + \Delta \leq \int p(\xi_T; \pi_T) R(\xi_T) d\xi_T \quad (22b)
\]

Note that this is an alternative formulation of the relative entropy constrained problem defined by for example [36] and [37] amongst others. However here we flip the objective and constraint so to reveal the connection with Entropic Inference which we argue is the underlying principle. On account of the method of Lagrangian multipliers either problem formulation ultimately amounts to the same mathematical problem after appropriate scaling of the Lagrangian multipliers or cost function.
When we adopt appropriate scaling of $R$ and $\Delta$ so that the Lagrangian multiplier has value 1 we retrieve the earlier problem from (16)

$$
\mathbb{D} \left[ p(\xi_T; \pi_T) \parallel p(\xi_T^0; \pi_T) \right] + \int p(\xi_T; \pi_T) R(\xi_T) d\xi_T
= \int p(\xi_T; \pi_T) \log \frac{p(\xi_T; \pi_T^0) e^{-R(\xi_T)}}{p(\xi_T^0; \pi_T) e^{-R(\xi_T)}} d\xi_T
= \mathbb{D} \left[ p(\xi_T; \pi_T) \parallel p(\xi_T^0; \pi_T) e^{-R(\xi_T)} \right]
$$

(23)

### 3.2 Regularized Risk Sensitive Optimal Control

In this section we retake from problem (4). Before we submit it to a similar analysis as was conducted above, we set out to seek an analogy with Bayesian estimation. This analogy will allow us to approach the problem intuitively. For a prior on Bayesian estimation, we refer to Appendix B.

**Figure 2.** Graphical representation of a Controlled Hidden Markov Model.

**3.2.1 Equivalent Bayesian estimation problem** To establish a rigorous connection with the Bayesian estimation framework, first we extend the CMC model, as it was introduced in section 2.1, with a probabilistic measurement model or so-called emission probability $p(z_t|\xi_t)$, see Figure 2. We refer to the corresponding probabilistic graph model as a Controlled Hidden Markov Model (CHMM).

The measurements are considered independent of any history of the trajectory apart from the present state and action. The probability of a measurement sequence, $\zeta = \{z_0, \ldots, z_T\}$, conditioned on a trajectory, $\xi_T$, can then be defined as

$$p(\zeta_T|\xi_T) = \prod_{t=0}^{T} p(z_t|\xi_t)$$

and the joint model parameterised by a policy sequence $\pi_T$ is then defined as

$$p(\xi_T, \zeta_T; \pi_T) = p(x_0)p(z_T|x_T)\prod_{t=0}^{T-1} p(z_{t+1}|x_{t+1})p(x_{t+1}|\xi_t)p(\xi_t)\pi_t(\xi_t)$$

(25)

Clearly this model shares its main structure with that of a classical Hidden Markov Model (CHMM). Application of BI procedures on this type of model are well studied and are easily extended to the case of CHMMs, see Appendix B. Two important examples are the Bayesian filtering problem where we seek the conditional

+ In fact, the latent action sequence could be marginalised out of the model, adopting closed loop transition and emission probabilities, $p(x_{t+1}|x_t; \pi_t) = \int p(x_{t+1}|\xi_t)\pi_t(\xi_t)d\xi_t$, and, $p(z_t|x_t; \pi_t) = \int p(z_t|\xi_t)\pi_t(\xi_t)d\xi_t$, respectively. A process referred to as lumping.
To establish a connection with control the following emission model is proposed \cite{38,8,39}. We introduce an auxiliary variable $z_t$ that we assume to have adopted some artificial and arbitrary value with probability proportional to $e^{-r_t(\xi_t)}$. The associated joint model is also given.

$$p(z_t | \xi_t) \propto e^{-r_t(\xi_t)}$$  \hspace{1cm} (26)$$

$$p(z_T | \xi_T) \propto e^{-R(\xi_T)}$$

Now let us reconsider the MLE problem

$$\max_{\pi_T} \log p(z_T; \pi_T)$$  \hspace{1cm} (27)$$

Using the substitutions proposed above one verifies that this problem is equivalent to the risk sensitive optimal control problem in (4)

$$p(z_T; \pi_T) = \int p(z_T | \xi_T) p(\xi_T; \pi_T) d\xi_T \propto \int p(\xi_T; \pi_T)e^{-R(\xi_T)} d\xi_T$$  \hspace{1cm} (28)$$

This observation is fundamental in that it establishes that there is no technical difference between the Bayesian argument developed here and the underlying RSOC problem. So we may treat this problem further by addressing it as a standard Bayesian estimation problem rather than an optimal control problem and treat it in the same way as we have treated \cite{2} in \cite{15}.

We further our analysis by decomposing the problem into two subproblems.

$$\log p(z_T; \pi_T) = \int q(\xi_T) \log \frac{p(z_T | \xi_T) p(\xi_T; \pi_T)}{q(\xi_T)} d\xi_T + \int q(\xi_T) \log \frac{q(\xi_T)}{p(\xi_T | z_T; \pi_T)} d\xi_T$$  \hspace{1cm} (29)$$

As opposed to developments in section 3.1 here we introduce a generic inference density, $q(\xi_T)$, similar to that in section 2.2.3 Application of the EM algorithm then boils down to the following steps.

**E-step** Minimise the error for $q$ keeping $\pi_T$ fixed.

**M-step** Maximise the ELBO for $\pi_T$ keeping $q$ fixed.

In this case the solution of the E-step is trivial. Specifically, the relative entropy is rendered minimal for $q(\xi_T) \equiv p(\xi_T | z_T; \pi_T)$. Since we will iterate this solution, at this point it is better to replace $q(\xi_T)$ with the more familiar expression $p(\xi_T | z_T; \rho_T)$ where $\rho_T$ is substituted for the now old or previous control belief sequence. This substitution proposed earlier now makes it less confusing that in the M-step we are trying to find an updated version of $\pi_T$ than the version that is contained in the expression for the inference density, $q(\xi_T)$.

The question here thus reduces to solving the M-step. Note that the emission probability $p(z_T | \xi_T)$ can be omitted without altering the solution of the ELBO.

$$\pi_T^* = \arg \min_{\pi_T} \mathbb{D} \left[p(\xi_T | z_T; \rho_T) \parallel p(\xi_T; \pi_T)\right]$$  \hspace{1cm} (30)$$
3.2.2 Explicit solution of the M-step

Let us first further reduce problem (30)

\[ \arg \min_{\pi_T} \mathbb{D} \left[ p(\xi_T|x_T; \rho_T) \parallel p(\xi_T|x_T; \pi_T) \right] = \arg \min_{\pi} \int p(\xi_T|x_T; \rho_T) \log \frac{p(\xi_T|x_T; \rho_T)}{p(\xi_T|x_T; \pi_T)} d\xi_T \]

\[ = \arg \min_{\pi} \int p(\xi_T|x_T; \rho_T) \sum_{t=0}^{T-1} \log \pi_t(\xi_t) d\xi_T \]

\[ = \arg \min_{\pi} \sum_{t=0}^{T-1} \int p(\xi_t|x_T; \rho_T) \log \pi_t(\xi_t) d\xi_t \]

This decomposition illustrates that as opposed to problem (16), the present optimisation problem is not subject to an optimal substructure and can be treated independently for each separate policy. Taking into account the normalisation condition it follows that

\[ \pi_t^*(\xi_t) = \frac{p(\xi_t|x_T; \rho_T)}{p(x_t|x_T; \rho_T)} = p(u_t|x_t, z_T; \rho_T) \]  

The solution of (30) is thus equivalent to the probability of \( u_t \) parameterised by the prior policy sequence, \( \rho_T \), and conditioned onto the measurements, \( z_T \). Before we venture into further technicalities we can submit the densities, \( \pi_T^* \), to a preliminary assessment.

Since we conditioned on the state \( x_t \), it follows that this density must be equivalent to \( p(u_t|x_t, z_T; \pi_t) \) rather than \( p(u_t|x_t, z_T; \rho_t) \). This is a direct result of the Markov property, since no more information about the input at time instant \( t \) can be contained in the older measurements, \( z_{t-1} \), than is already contained by the state, \( x_t \), itself. This resonates with the common sense that once we have arrived at some state, \( x_t \), we can only hope to reproduce the measurements, \( z_t \), but can no longer affect any of the preceding measurements, \( z_{t-1} \). So again, some form of optimal substructure is present although here it is not a property of the optimisation problem but rather a property of the conditional probability \( p(u_t|x_t, z_T; \rho_T) \).

Now using Bayes’ rule, \( \pi_t^* \) can be decomposed as

\[ p(u_t|x_t, z_t; \pi_t) = \rho_t(\xi_t) \frac{p(z_t|x_t; \pi_{t+1})}{p(z_t|x_t; \pi_t)} \]  

which reduces the problem to finding efficient expressions for the probabilities \( p(z_t|x_t; \pi_t) \) and \( p(z_t|x_t; \pi_{t+1}) \). The latter is a generalisation of the backward filtering densities.

We have that

\[ p(z_t|x_t; \pi_t) = \int \rho_t(\xi_t) p(z_t|x_t; \pi_{t+1}) d\xi_t = \mathbb{E}_{\rho_T(\xi_t)} \left[ p(z_t|x_t; \pi_{t+1}) \right] \]

These backward filtering densities adhere to a backward recursive expression.

\[ p(z_t|x_t; \pi_{t+1}) = p(z_t|x_t) \int p(x_{t+1}|x_t) p(z_{t+1}|x_{t+1}; \pi_{t+1}) dx_{t+1} \]

\[ = p(z_t|x_t) \mathbb{E}_{p(x_{t+1}|x_t)} \left[ p(z_{t+1}|x_{t+1}; \pi_{t+1}) \right] \]

Then for, fixed \( \rho_T \) (and \( z_T \)), we can recycle the definition of \( \pi_T^* \) and define the functions

\[ Q_t^*(\xi_t) = -\log p(z_t|x_t; \pi_{t+1}) \]

\[ V_t^*(x_t) = -\log p(z_t|x_t; \pi_t) \]
By definition then it follows that
\[
\pi_t^\star(\xi_t) = \rho_t(\xi_t) \frac{\exp(-Q_t^\star(\xi_t))}{\exp(-V_t^\star(x_t))}
\]
\[
V_t^\star(x_t) = -\log \mathbb{E}_{p(u_t|x_t;\pi_t)}[\exp(-Q_t^\star(\xi_t))]
\]
\[
Q_t^\star(\xi_t) = r_t(\xi_t) - \log \mathbb{E}_{p(x_{t+1}^\star|\xi_t)}[\exp(-V_{t+1}^\star(x_{t+1}))]
\]

Again, we have found a backward recursive procedure to derive a sequence of policies, \(\pi_T^\star\), initialising \(V_T^\star(x_T)\) as \(\exp(-r_T(x_T))\). Then by practising the substitution \(\rho_T \leftarrow \pi_T^\star\) and repeated application of the equations in (37), now a program is specified to solve the variational optimisation problem in (4). On first notice the solutions of the problems (16) and (30) are equivalent. The difference lies in the definition of the functions, \(Q_T^\star\), and \(Q_T^\star\).

3.2.3 Alternative solution of (30) Before we discuss the result from this section in greater detail, we provide here an alternative derivation of the result in (37) and hence an alternative solution strategy tailored to problem (30). Although the solution strategy that was based on the technical equivalence with the Bayesian estimation paradigm is elegant, it may remain unconvincing to some since we rely on the Markovian properties of the model to simplify the calculations.

Consider therefore the following problem. Note that this problem is the reciprocal M-projection of, \(p(\xi_T; \rho_T) \exp(-R(\xi_T))\), onto the density space spanned by, \(p(\xi_T; \pi_T)\), of the I-projection in equation (16). We leave it to the reader to verify that this problem is indeed equivalent to (30) and that therefore the Bayesian equivalence is strictly unnecessary.

\[
\pi_T^\star = \arg \min_{\pi_T \in \mathcal{P}} \mathbb{D} \left[ p(\xi_T; \rho_T) e^{-R(\xi_T)} \parallel p(\xi_T; \pi_T) \right] \tag{38}
\]

Similar to the derivation laid out in (31) we have that
\[
\pi_T^\star(\xi_t) = \frac{v_t(\xi_t)}{\int v_t(\xi_t)d\xi_t} \tag{39}
\]
where
\[
v_t(\xi_t) = \int p(\xi_{t+1}; \rho_T) e^{-R(\xi_{t+1})} d\xi_{t+1} \tag{40}
\]

The function \(v_t\) can be decomposed as follows
\[
v_t(\xi_t)
\approx \rho_t(\xi_t) e^{-r_t(\xi_t)} \int p(\xi_{t-1}; x_t; \rho_{t-1}) e^{-R(\xi_{t-1})} d\xi_{t-1} p(\bar{\xi}_{t+1}|\xi_t; \bar{\pi}_{t+1}) e^{-R(\bar{\xi}_{t+1})} d\bar{\xi}_{t+1}
\]
\[
\times \rho_t(\xi_t) e^{-r_t(\xi_t)} w_t(x_t) \int p(\bar{\xi}_{t+1}|\xi_t; \bar{\pi}_{t+1}) e^{-R(\bar{\xi}_{t+1})} d\bar{\xi}_{t+1}
\]

Now recall that the policy is established by normalisation on \(u_t\), which implies that the factor, \(w(x_t)\), occurs in both nominator and denominator and does not affect the result by consequence. Then one easily retrieves the recursive expressions in (37).
On Optimal Control and Expectation-Maximisation

Table 1. Overview of the iterative programs suggested by (21) and (37).

| \( \pi_{t+1}^{g}(\xi_t) \) | SOC | RSOC |
|-----------------|-----|------|
| \( \pi_{t}^{g}(\xi_t) \) | \( \exp(-Q_{t+1}^{g}(\xi_t)) \) | equivalent |
| \( V_{t+1}^{g}(x_t) \) | \( -\log \mathbb{E}_{p(u_t|x_t;\pi_{t}^{g})}[\exp(-Q_{t+1}^{g}(\xi_t))] \) | equivalent |
| \( Q_{t+1}^{g}(\xi_t) \) | \( r_t(\xi_t) + \mathbb{E}_{p(x_{t+1}|x_t)}[\exp(-V_{t+1}^{g+1}(x_{t+1}))] \) | \( r_t(\xi_t) - \log \mathbb{E}_{p(x_{t+1}|\pi_{t}^{g})}[\exp(-V_{t+1}^{g+1}(x_{t+1}))] \) |

4 Discussion

In the previous section we have treated the optimal control problems stated in (2) and (4). Our treatment is based on the EM algorithm. In either case an iterative program is suggested that converges to a local solution of the respective control problems.

In this section we explore some of the properties of these programs. We further establish connections with the field of Path Integral Control. Explicit evaluation of the backward recursive expressions in either program are however still restricted to a limited class of well-behaved optimal control problems. We give an overview of those in the second to last paragraph. Finally, we discuss a heuristic that might be beneficial to improve the convergence of practical algorithms.

4.1 Iterating in policy belief space

Table 1 summarises the iterative programs that are suggested by (21) and (37). Comparison with the equations that govern the classical recursive solution of the problems in (2) and (4) (see Appendix A) reveals that an interesting analogy is present. To recover the equations that govern the iterative probabilistic programs, the operator, \( -\log \mathbb{E}_{p(u_t|x_t;\pi_{t}^{g})}[\exp(-\cdot)] \), over controls given the prior policy model, is substituted for the minimisation operator, \( \min_{u_t}(\cdot) \), in the equations that govern the classic solution. Ergo the classic direct minimisation operator is replaces by an iterative probabilistic inference operator, that instead of spawning the deterministic solution immediately, generates a convergent sequence of densities.

Indeed we have that by definition the elements of the sequence, \( \pi_{T}^{g} \), are in the set of densities \( \mathcal{P} \). Therefore it is tempting to interpret the functions in \( \pi_{T}^{g} \) as probability densities. From a technical point of view such an interpretation is irrelevant. Though here we argue that one may interpret each iterate policy as a set of belief functions that express our uncertainty about the underlying deterministic solution. Put differently, the sequences are an expression of our epistemic uncertainty. This point of view corresponds with the line of thought developed in section 3.1.2, where we provided an alternative derivation of the results that are tailored to SOC. Provided the analogy with Bayesian estimation, such an interpretation is even easier to assign to the results tailored to the RSOC case.

At this point we also would like to bring to the attention, that the problems, (16), (I-projection) and (22), (M-projection), determine each others reciprocal information-theoretic projection of the density, \( p(\xi_T;\rho_T)e^{-R(\xi_T)} \), onto the space spanned by the parameterise trajectory density, \( p(\xi_T;\mathcal{X}_T) \). This observations hints at an intimate relation between information-theory and optimal control.
4.2 Monotonicity and convergence properties

Until now we have voiced little concern whether the fixed point iterations associated to the programs defined by (21) and (37) are convergent and to what stationary point. To put it more formally, will the limit policy, \( \lim_{g \to \infty} \pi^g_T \), coincide with the actual optimum? The answer is suggested by the convergence properties of the EM algorithm. As has been shown by [24, 32], in the EM algorithm converges monotonically to a saddle point or a local maximum of the likelihood function (here a local minimum of the cost function) and so to the solution of the original inference problem if the problem is sufficiently well-behaved. In practice this usually boils down to a local minimum which is satisfactory for many applications. Hence this monotonicity property translates to the iterative programs studied here, provided that the the initial policies, \( \pi^0_T \), covers the support of the true solutions, \( \pi_T \).

Lemma 1. Consider a sequence defined as \( \{\pi^g_T\}_{g \in \mathbb{N}} \), that is initialised uniformly on \( \mathcal{U} \), and whose elements are found by substituting \( \pi^g_T \leftarrow \pi^{g+1}_T \) in (21) or either by substituting \( \pi^g_T \leftarrow \pi^{g+1}_T \) in (37). This sequence converges to a saddle point or a local solution of either problem (2) or (4).

Whether a practical algorithm will converge to the optimal solution still depends on the characteristics of said algorithm and the approximations that are made.

Recalling the discussion in section 3.1.2 we also note that the converge behaviour of the iterative program tailored to the SOC problem, can be affected by changing the value of the Lagrangian multiplier, \( \lambda \), without affecting the associated optimal control problem; that, whilst we cannot influence the convergence behaviour of the iterative program tailored to the RSOC problem. The situation is similar to their regular counterparts where the cost function can be scaled with SOC but not RSOC.

4.3 The limit behaviour of (37)

In the regular case, the RSOC problem is sometimes viewed as a generalisation of the SOC problem. Recall indeed that for the limit \( \sigma \to 0 \), the RSOC framework collapses on the SOC framework. Furthermore, for deterministic dynamics both problem collapse on the same deterministic problem. Let us here verify whether such characteristic behaviour is maintained in the present context.

As we have already commented on earlier, the iterative programs suggested by (21) and (37) are equivalent except for the definition of the associated functions, \( Q^\bullet_t \) and \( Q^\triangle_t \). Hence we should only verify that either \( Q^\bullet_t \) and \( Q^\bullet_t \) are equivalent when the transition model is limited to the set of deterministic functions. One easily verifies this is indeed the case, as for deterministic dynamics the following identity holds

\[
Q^\bullet_t(\xi_t) = Q^\bullet_t(\xi_t) = Q_t(\xi_t) = r_t(\xi_t) + V_{t+1}(f_t(\xi_t)) \quad (42)
\]

Progressing down that arc we can now also study the limit behaviour of (37) for \( \sigma \to 0 \). Consider therefore the auxiliary functions \( V'_t \) and \( Q'_t \) defined as

\[
Q'_t(:; r_t, \sigma) = -\frac{1}{\sigma}Q^\bullet_t(:; -\sigma r_t) \\
V'_t(:; r_t, \sigma) = -\frac{1}{\sigma}V^\bullet_t(:; -\sigma r_t) \quad (43)
\]

One can verify the following limits. Note that the limits needs to be evaluated recursively, starting with \( V^\bullet_T(x_T) = V^\bullet(x_T) = r_T(x_T) \) and relying on the classical limit \( \lim_{\sigma \to 0} -\frac{1}{\sigma} \log \mathbb{E}[\exp(-\sigma f)] = \mathbb{E}[f], f \geq 0 \) which may be verified using L'Hôpital's rule.
\[ Q^*_t(:; r_t) = \lim_{\sigma \to 0} Q'_t(:; r_t, \sigma) \]
\[ V^*_t(:; r_t) = \lim_{\sigma \to 0} V'_t(:; r_t, \sigma) \]  
(44)

As could be anticipated, the characteristic relation between RSOC and SOC is preserved regardless our treatment with the EM algorithm.

4.4 Relations with Path Integral Control

Another interesting observation can be made regarding the function, \( V^*_t \). Considering the expression developed in either section (3.2.2) or (3.2.3), we observe the following
\[ V^*_t(x_t) = -\log E_{p(\xi_t|x_t; \rho_t)} \left[ e^{-R(\xi_t)} \right] \]  
(45)

and therefore
\[ \pi^*_t(\xi_t) \propto \rho_t(\xi_t) E_{p(\xi_{t+1}|\xi_t; \rho_t)} \left[ e^{-R(\xi_t)} \right] \]  
(46)

This observation is very practical. It illustrates that the value function can be computed as a Path Integral over the trajectory density parameterised by the prior policies. Consequently we could estimate \( V^*_t \) by sampling from \( p(\xi_T|x_t; \rho_t) \). This property only holds for the RSOC case. But since the SOC and RSOC problems are equivalent for deterministic dynamics, the property extends generally to deterministic optimal control. The existence of a path integral solution in the context of deterministic Entropic Optimal Control was also described by [40].

There is another optimal control framework that amounts to such conditions, better known as Linearly Solvable Optimal Control or Kullback-Leibler control (also see section 5). As detailed in the introduction the origins of this framework root back to continuous time optimal control however this is out of scope here. We will discuss the framework briefly in discrete time [41, 42, 43]. In discrete time the LSOC framework is achieved by lifting the regular SOC entirely to the state space [44]. To link this to control we assume that we can directly control the transition probability \( p(x_{t+1}|x_t; \pi_t) \). This can be achieved by e.g.
\[ p(x_{t+1}|x_t; \pi_t) = \int p(x_{t+1}|x_t, u_t)\pi_t(u_t|x_t)dx \]  
(47)

and assuming that for every \( p(x_{t+1}|x_t; \pi_t) \) we can find some \( \pi_t \) that achieves it.

The LSOC problem in discrete time can then be stated as
\[ \min_{\pi} \mathbb{E}_{p(\xi_T; \pi_T)} \left[ R(\xi_T) + \log \frac{p(\xi_T; \pi_T)}{p(\xi_T; \rho_T)} \right] \]  
(48)

where usually \( \rho_T = \mathbb{U} \).

Provided with the theory developed in this section it is easily verified that
\[ p(x_{t+1}|x_t; \pi_t) \propto p(x_{t+1}|x_t; \rho_t)e^{-V_{t+1}(x_{t+1})} \]  
(49)

and
\[ V_t(x_t) = r_t(x_t) - \log \mathbb{E}_{p(x_{t+1}|x_t; \pi_t)} \left[ \exp(-V_{t+1}(x_{t+1})) \right] \]
\[ = -\log \mathbb{E}_{p(\xi_{t+1}|x_t; \mathbb{U})} \left[ e^{-R(\xi_t)} \right] \]  
(50)
Clearly this result is generalised to RSOC problems, providing a theoretical basis to the wide variety of algorithms based on Path Integral Control amongst which most notably the collection of Model Based Path Integral control methods described in section 5. Furthermore this also formulates an answer posed by [39] who investigated the relation between Linearly Solvable Optimal Control and message passing in the context of Control as Inference.

4.5 Specialisation to linear-Gaussian models and Quadratic costs

The results from (21) and (37) can only be solved explicitly for a limited class of SOC and RSOC problems. Discrete space problems are one example. In this second to last paragraph, we specialise the results to the Linear Quadratic Regulator (LQR) and the Linear Exponential Quadratic Regulator (LEQR) which are examples of the SOC and RSOC problem specialised to linear-Gaussian dynamics, \( p(x_{t+1} | \xi_t) = \mathcal{N}(x_{t+1} | F_{\xi,t} \xi_t + f_t, P_t) \), and Quadratic costs, \( r_t(\xi_t) = \frac{1}{2} \xi_t^T R_{\xi,t} \xi_t + \xi_t^T R_{\xi,t} + \ldots \) and \( r_T(x_T) = \frac{1}{2} \xi_T^T R_{xx,T} x_T + x_T^T R_{x,T} + \ldots \).

Since the results in (21) and (37) differ only in the expression for \( Q^*_{t} \) and \( Q^*_t \), we can develop shared equations for \( \pi^*_t \) and \( V^*_t \) where \(* \in \{\bullet, \triangle \} \). In this context it is reasonable to assume that the prior policy densities are linear-Gaussian, \( p_t(\xi_t) = \mathcal{N}(u_t | K_{t}x_t + k_t, \Sigma_t) \), that both \( Q^*_t \) and \( V^*_t \) will be quadratic in their arguments and that the posterior policies, \( \pi^*_t(\xi_t) = \mathcal{N}(u_t | K_{t}x_t + k^*_t, \Sigma^*_t) \), are also linear-Gaussian in the state. Here \( \times \) indicates unspecified (or redundant) information and \( \cdot \) symmetric matrix entries.

\[
V^*_t(x_t) = \frac{1}{2} \begin{bmatrix} 1 \\ x_t \end{bmatrix}^T \begin{bmatrix} V^*_{x,t}^T & 1 \\ \cdot & V^*_{xx,t} \end{bmatrix} \begin{bmatrix} 1 \\ x_t \end{bmatrix}
\]

\[
Q^*_t(\xi_t) = \frac{1}{2} \begin{bmatrix} 1 \\ \xi_t \end{bmatrix}^T \begin{bmatrix} Q^*_{\xi,t}^T & 1 \\ \cdot & Q^*_{\xi\xi,t} \end{bmatrix} \begin{bmatrix} 1 \\ \xi_t \end{bmatrix}
\]

Since \( \pi^*_t(\xi_t) \propto p_t(\xi_t) \exp(-Q^*_t(\xi_t)) \) one verifies that

\[
k^*_t = \Sigma^*_t (\Sigma^{-1}_t k_t - Q^* u_t, t)
\]

\[
K^*_t = \Sigma^*_t (\Sigma^{-1}_t K_t - Q^*_{x,t})
\]

\[
\Sigma^*_t = (\Sigma^{-1}_t + Q^*_{uu,t})^{-1}
\]

As \( \exp(-V^*_t(x_t))\pi^*_t(\xi_t) = p_t(\xi_t) \exp(-Q^*_t(\xi_t)) \) for any \( u_t \) including 0 we have

\[
V^*_{x,t} = Q^*_{x,t} + K^*_t \Sigma^{-1}_t k_t - K^*_t \Sigma^{-1}_t k^*_t
\]

\[
V^*_{xx,t} = Q^*_{xx,t} + K^*_t \Sigma^{-1}_t K_t - K^*_t \Sigma^{-1}_t K^*_t
\]

Finally we can evaluate the specific expressions of the functions \( Q^*_t \) and \( Q^*_t \).

**Entropic LQR** The entropy regularised sibling of the LQR (E-LQR) equals

\[
Q^*_{\xi,t} = R_{\xi,t} + F^*_{\xi,t} (V^*_{xx,t+1} f_t + V^*_{x,t})
\]

\[
Q^*_\xi,t = R_{\xi,t} + F^*_{\xi,t} V^*_{xx,t+1} F_{\xi,t}
\]

Note that the stochasticity of the dynamics, namely \( P_t \), only affects the constant terms of the quadratic functions, \( Q^*_t \), and, \( V^*_t \), (not given here), as is the case for the standard LQR. The policy itself is only affected by the linear and quadratic terms. It follows that the solution is equivalent for stochastic as well as for deterministic systems.
Entropic LEQR The entropy regularised sibling of the LEQR (E-LEQR) equals

\[
Q^\xi_{\xi,t} = R_{\xi,t} + F_{\xi,t}^T (V^\xi_{\xi,t+1}^{-1} + P_t)^{-1} \left( V^\xi_{\xi,t+1}^{-1} V^\xi_{\xi,t} + f_t \right)
\]

\[
Q^\xi_{\xi\xi,t} = R_{\xi\xi,t} + F_{\xi,t}^T (V^\xi_{\xi,t+1}^{-1} + P_t)^{-1} F_{\xi,t}
\]

As opposed to the solution detailed above, here the update does depend on the stochasticity of the dynamics. This agrees with the underlying theory of risk sensitive optimal control developed by Jacobson specifically because the regular LQR and E-LQR do not possess this property. Despite appearing benign this observation also nicely demonstrates another difference with the (E-)LQR. It imposes a proportionality between the value function, \(V^\xi_t\), and the stochasticity of the system, \(P_t\). It follows that by scaling the cost function, \(r_t\), the proportionality is altered between \(V^\xi_t\) and \(P_t\) and thus we also alter the underlying RSOC problem. This was not the case with the E-LQR. Following our discussion in section 4.2 in the context of the E-LQR, scaling of the cost function, \(r_t\), is equivalent to scaling of the corresponding Lagrangian multiplier. So reflecting on the iterative aspect of the programs, we would influence the convergence behaviour of the iteration not the properties of the underlying SOC problem. This verifies our earlier remark that as opposed to the convergence behaviour of the E-LQR, the convergence behaviour of the E-LEQR can not be influenced.

4.6 Prevention of distributional collapse

Although the theoretical convergence of described iterative programs is guaranteed on account of the characteristics of the EM algorithm, in practice it could occur that the procedure converges prematurely. This means that the iterated policy beliefs collapse onto a Dirac delta function that does not coincide with the underlying deterministic solution. In order to remedy this characteristic we can try to adapt the underlying inference procedures so that the density sequence simply does not converge to the Dirac delta function, even in the limit.

Let us discuss the stochastic optimal control problem first. Then assuming that we initialise the policy belief sequence with the uniform distribution, modelling the situation where we are completely uncertain about the true solution, the idea is to slightly reset the solution every iteration, solving a superposition of the iterate and the initial problem. Therefore we retake from problem (22) and extend the objective as the weighted sum over the two objectives mentioned. The degree with which the solution is reset is governed by the parameter \(0 < \alpha \leq 1\).

\[
\min_{\xi_T} \alpha \mathbb{D} \left[ p(\xi_T; \pi_T) \parallel p(\xi_T; \pi_T^\alpha) \right] + (1 - \alpha) \mathbb{D} \left[ p(\xi_T; \pi_T) \parallel p(\xi_T; \pi_T^\alpha) \right] \quad (56a)
\]

subject to

\[
\int p(\xi_T; \pi_T) R(\xi_T)d\xi_T + \Delta \leq \int p(\xi_T; \pi_T^\alpha) R(\xi_T)d\xi_T \quad (56b)
\]

Note that \(1 - \alpha\) can also be interpreted as a Lagrangian multiplier associated to the inequality constraint \(\mathbb{D} \left[ p(\xi_T; \pi_T) \parallel p(\xi_T; \pi_T^\alpha) \right] > \Delta\) imposed on optimisation problem (22). This idea is in that sense closely related to the works of [45, 46]. Though by setting for the Lagrangian multipliers instead of meeting hard constraints, the influence of this procedure manifests more clearly in the associated policy update.
To see this one can verify that, similar to the previous derivation in (22), the solution is now governed by the following I-projection.

$$\min_{\pi} \mathbb{D} \left[ p(\xi_T; \pi_T) \| p(\xi_T; \pi_T^g, \alpha) e^{-R(\xi_T)} \right]$$

(57)

One agrees that this amounts to a diffused version of problem (16) given that the previous belief is diffused with a factor $0 < \alpha \leq 1$ effectively keeping the inferred belief from collapsing onto the Dirac-delta distribution. It is also easily seen that this problem is now equivalent to

$$\min_{\pi} \mathbb{D} \left[ p(\xi_T; \pi_T) \| p(\xi_T; \pi_T^g) e^{-R^g(\xi_T; \alpha)} \right]$$

(58)

where

$$R^g(\xi_T; \alpha) = r_T(x_T) + \sum_{t=0}^{T-1} r^g_t(\xi_t; \alpha)$$

(59)

$$r^g_t(\xi_t; \alpha) = r_t(\xi_t) + (1 - \alpha) \log \pi^g_t(\xi_t)$$

so that the same effect can be achieved by substituting the cost rate $r^g_t$ for $r_t$ whilst retaining the updates for $\pi^g_t$ and $V^g_t$.

A similar treatment for the RSOC case can not be achieved from first principles because of the mechanics of the problem. However a similar result can still be achieved by substituting $\pi_T^g, \alpha$ for $\pi_T^g$ in the M-projection (38) or by substituting the cost model introduced above. So although we then lack a motivation founded on first principles, we can in fact adopt similar diffused updates.

5 Related work

Since the conception of modern system theory, the history of control and estimation have been intimately related, cultivating cautious aspirations to pursue dualities and equivalences early on. In the tradition of other summaries of this kind, reminiscing about the historical roots of optimal control and estimation, we honour the seminal work of Rudolf Kalman [47, 48] and the duality between the Linear Quadratic Regulator and the Kalman Filter. Unfortunately it proofed difficult to generalise such duality beyond a linear setting and not for a lack of trying [49, 50, 51].

It is fairly safe to state that a serious system-theoretic interest in inferring the optimal policy traces back to the discovery of Linearly Solvable Optimal Control (LSOC) by Kappen [52], followed by that of Linearly Solvable Markov Decision Processes by Todorov [41], which was later extended by [43] and [35]. These frameworks refer to a peculiar subclass of SOC problems where the optimal policy can be expressed as a path integral over future passive or uncontrolled trajectories, in the sense of Feynman and Kac [53, 54]. Put more explicitly, the optimal policy can be expressed in terms of a conditional expectation over the uncontrolled path distribution. Contemporary, the framework is often referred to as Path Integral Control (PIC) and derived methods as PIC methods. Soon it was explored how these results could be leveraged to design actual policies by evaluating the integral [55] or by solving an eigenvalue problem [56].

Following earlier work from about the same period as LSOC [57], in 2009 Toussaint [58, 38] pioneered the idea to formulate a probabilistic graphical model whose most likely instance coincided approximately with the optimal one and used
it as a theoretical basis to derive a gradient based trajectory optimisation algorithm based on the concept of message passing [24]. Toussaint was therefore the first to exploit the similarity between the graphical models underlying the Bayesian estimation framework and optimal control. This approach is one instance of Cal, often denoted as Inference for Control or Inference to Control (I2C).

Another field that clearly showcases interest in both control and inference is Reinforcement Learning (RL). Contemporary the problem of RL is cast as a stochastic optimal control problem. The main ambition in RL is to learn the optimal policy simply by eliciting and observing system-environment interactions. As noted by Peters the use of data as a means to learn or extract the optimal policy then relies on tricks rather than a factual build-in inference mechanism [59]. Put differently, learning is possible only due to the fact that the cost we try to minimise contains an expectation but not necessarily expresses the policy itself in terms of one. Take for example policy search methods that simply parameterise the policy and rely on the stochasticity of the dynamics for sufficient exploration. Until now LSOC is the only framework that expresses the policy itself as an expectation and that does not require any tricks to do so. Nevertheless also here the source of stochasticity roots back to the system.

In an attempt to address such principle shortcomings, the RL community proposed (indirectly) information-theoretic connections to the effect of regularising SOC problems by introduction of information-theoretic measures such as the relative and differential entropy. Peters [36] explored regularising the objective with a relative entropy term inspired by the superior performance of natural over vanilla gradients in policy search methods [60–61]. The work on relative entropic RL culminated into the state-of-the-art trust region policy optimisation method [37] and closely related proximal policy algorithm [62]. Second, say that we try to infer a cost model from (human) demonstrations, then the standard stochastic optimal control framework makes for a poor model, as it can not explain the (minor) suboptimalities that might be present in demonstrations. To address this issue, Ziebart [63] introduced a differential entropy term in the control objective and described the solution of the Entropy Regularised Linear Quadratic Gaussian Regulator. Interestingly enough, the variational solution to these regularised problems are stochastic policies, or to be more precise, the optimal policy is given by a density rather than a deterministic function. One quickly argued that these ideas might offer an answer to the problem of exploration in RL. This stimulated the development of Maximum Entropy or soft RL algorithms [64–66]. Meanwhile work was done attempting to unify the ideas of Toussaint with the relative entropy framework and LSOC [6].

In parallel the work on LSOC and PIC also proceeded. In 2010, Theodorou devised the first RL policy search algorithm from the theory of LSOC [67, 68]. The work of [69] hinted at a deeper connection with stochastic optimisation techniques inspiring the first theoretical investigations on dualities between information theory and stochastic optimal control [70, 71] and refuelling work on combining LSOC with the ideas from RL on entropy regularisation [72, 73, 74, 75]. Finally during the second part of the previous decade a series of model based algorithms were proposed that make use of the fact that in the context of LSOC the policy can be expressed as a path integral to update the control based on sampled trajectories [76–78, 81]. Common applications are trajectory optimisation [79, 80], Model Predictive Control [81, 82], and Reinforcement Learning [84]. More recently a generalised PIC problem was proposed that unified LSOC with the Relative Entropy framework [35]. Noteworthy is the underlying principle of Entropic Optimisation and the first mentioning of the
Entropic Optimal Control (EOC) framework. In [40] this view was generalised showing that for deterministic systems, the EOC framework also gives rise to path integral expressions for the optimal policy completing the unification between PIC and EOC.

Finally we mention a series of work that directly derive from the view on optimisation pioneered by Toussaint [38], here referred to as I2C. This idea was embraced by a number of works in Reinforcement Learning to derive a series of RL approaches such as the work of [53, 7, 63] amongst others. The latest RL algorithm that is inspired by the inference for control paradigm is Maximum a posteriori Policy Optimisation (MPO) [86, 87, 88]. Finally we can mention the work of Watson et al. who recently revived the pioneering work by Toussaint on I2C and message passing algorithms [89, 90, 91] and attempted unifications with LSOC [39].

6 An Outlook Towards Algorithms

Building on the theoretical results featured in sections 3 and 4 here we discuss a number of possibilities to mould these ideas into new or improved practical algorithms. As noted in the introduction, algorithmic development based on inference principles in control, has remained an activity endeavoured mostly by the RL community, and less so by the control and robotics community. The efforts of the RL community usually focuses on global solution methods and has produced a variety of RL algorithms tailored to global policy optimisation. The control and robotics community, on the other hand, exhibit particular interest in trajectory optimisation. That is why we focus here on trajectory optimisation algorithms (or thus local methods). In the present settings, this means that we seek for an expression of the policy that is only valid in the proximity of a narrow trajectory density. This condition is satisfied when the initial state model, \( p(x_0) \), is reasonably narrow or collapses on a deterministic value.

Further, we treat here trajectory optimisation in settings where the iLQR algorithm would apply. We refer readers unfamiliar with that setting to some of the standard works [14, 13, 92, 15, 21]. We also limit our exposition to deterministic system dynamics, in part because then the SOC and RSOC settings amount to the same expressions. Second this allows us to narrow down the scope to sketching the main ideas, omitting all too technical details that would extend to the treatment of stochastic dynamics. Finally in all of the algorithms described here we can use the auxiliary control cost, \( r^g_T \), as defined in section 4.6. This aspect will not be emphasised throughout and can be considered as a silent feature of all documented approaches.

Algorithm 1 Basic iLQR

1: \textbf{while} not converged \textbf{do}
2: \hspace{0.5cm} apply \( \tilde{\pi}_{i+1}^T \) to system to find \( \xi_{i+1}^T \) \hspace{0.5cm} \( \triangleright \) \textit{fup}
3: \hspace{0.5cm} solve a local problem approximation about \( \xi_{i+1}^T \) to find \( \hat{\xi}_{i+1}^T \) \hspace{0.5cm} \( \triangleright \) \textit{bwp}

As discussed in 2.1.3 the main idea of the iLQR algorithm is to iterate a local approximation of the true optimal policy sequence, \( \tilde{\pi}_i^T \), in proximity of some reference trajectory, \( \xi_i^T \). An essential characteristic of the algorithm is that it alternates a

* This means solving the stochastic optimal control problem for a global (deterministic) policy that can be used for any given initial state.
forward (fwp) and a backward pass (bwp). First the iterate policy is applied closed-loop to the system in order to spawn the new reference trajectory, $\xi_{i+1}$, $\Sigma_T$. In other words the present policy is evaluated. Second the new policy approximation, $\tilde{\pi}_{i+1}$, is determined according to some local approximation of the true optimal control problem about the new reference trajectory. In other words now the policy is updated. These two steps are repeated until some convergence conditions are met. The idea that we alternate between an evaluation and an optimisation step is fundamental characteristic of the iLQR algorithm. It is also implied that information about the cost landscape is only encoded in the solution during the backward pass and not during the forward pass. This is another important characteristic of the iLQR algorithm. The main steps of the algorithm are sketched in Algorithm 1. As the name suggests, a linear-quadratic approximation is used about the reference trajectory, $\xi_T$, so that the local policy approximation, $\tilde{\pi}_t = k_t^i + K_t^i x_t$, can be determined using the LQR solution. Typically this algorithm requires many numerical heuristics to stabilise the computations.

In that same spirit we limit ourselves to methods to finding local approximations of the global policy belief sequence $\pi_T$, in particular local Gaussian policies.

$$\pi_t(\xi_t) \approx \tilde{\pi}_t^g(\xi_t) = \mathcal{N}(u_t; k_t^g + K_t^g x_t, \Sigma_t^g)$$ (60)

Next we set out directions for two conceptual algorithmic families, respectively (ensemble) gradient based and sample based approaches.

6.1 Gradient based trajectory optimisation

This section focuses on methods that are gradient based or gradient like. We make a further distinction between two secondary algorithmic branches. (1) One branch is directly inspired by the iLQR algorithm. (2) The second branch explicitly exploits the connection between the EM treatment of RSOC and Bayesian estimation.

6.1.1 Iterated E-LQR (iELQR) The idea is fairly simple. We propose to extend the iLQR algorithm with a hyper iteration. In the hyper iterations we update the posterior policy, $\tilde{\pi}_T$, whilst in the main iterations we iterate the local approximation in the exact same way as the iLQR algorithm. Except here we use the E-LQR updates that conditions the policy update onto the prior policy belief sequence, $\tilde{\pi}_T$. The main idea is summarised in Algorithm 2. Here $j$ denotes the iteration counter for which the internal iteration has converged or for which we choose to interrupt it.

Some remarks are in place. Regardless whether the system itself is stochastic or deterministic, during the forward pass a path distribution is spawned on account of the policy beliefs. This means that we can not propagate the system in the same way as is done in the fully deterministic setting of the iLQR algorithm. Instead we will need to use a forward uncertainty propagation procedure. A straightforward choice is to propagate the uncertainty by means of a Taylor approximation similar to the Extended Kalman Filter (EKF, see Appendix B) but without the update step. Although here there is no update step, the closed-loop policy belief will stabilize the system about the local reference trajectory induced by that policy. Alternatively, we can also use a more sophisticated uncertainty propagation procedures such as the unscented transform used by the Unscented Kalman Filter (UKF, see Appendix B). As a by product either method produces locally linearised system dynamics as are later required by the backward pass. For the extended Kalman filter,
Algorithm 2 Iterated ELQR

1: while not converged do \( \triangleright \) hyper iterations
2: \hspace{0.5em} while not converged do \( \triangleright \) main iterations
3: \hspace{1em} apply \( \tilde{\pi}_T \) to system to find \( \xi^{i+1}_T \) \( \triangleright \) fwp
4: \hspace{1em} solve the LQER about \( \xi^{i+1}_T \) to find \( \tilde{\pi}_T^{i+1} \) conditioning on \( \tilde{x}_T \) \( \triangleright \) bwp
5: hyper update \( \tilde{\pi}_T^{i+1} \leftarrow \tilde{\pi}_T^j \)

Algorithm 3 Hyper iterated RTS

1: while not converged do \( \triangleright \) hyper iterations
2: \hspace{0.5em} run the iRTS smoother conditioning on \( \tilde{x}_T \) \( \triangleright \) main iterations
3: \hspace{1em} hyper update \( \tilde{\pi}_T^{i+1} \leftarrow \{p(u_t|x_t;\tilde{x}_T)\} \)

Ergo a Taylor approximation, this boils down eventually to exact gradients; when the unscented transform is used we practice an ensemble gradient approach.

The hypothesis is that the local optimisation landscape will be probed superiorly when using the unscented transform rather than the Taylor approximation. The use of the unscented transform to compute locally linear dynamics has already been used in the context of DDF\(^\dagger\)\cite{14}. Further comparison of the backward pass resulting from the implied iELQR (see section 4.5) with that of the iLQR algorithm (see Appendix A) illustrates that the inference based approaches eventually boil down to numerically regularised versions of the classic algorithm. It is emphasised that, 1) as opposed to iLQR, the iELQR algorithm is inherently iterative, and, 2) the regularisation is intrinsic and does not rely on any heuristics.

6.1.2 Hyper iterated RTS (HiRTS) The second idea is a bit more sophisticated. Following a similar line of thought as above, we can install a hyper iterative procedure and a secondary main iterative procedure. The hyper iteration has the same function as in the iELQR algorithm. In contrast, in the main iterative procedure we now propose to exploit the equivalence between the treatment of the RSOC problem using the EM algorithm and the corresponding Bayesian estimation problem. In the technical exposition accompanying section 3.2 we demonstrated that the optimal policy, \( \tilde{\pi}_T^{g+1} \), was given by the input conditioned on the state and artificial measurements, \( \tilde{x}_T \), and parameterised by the prior policy belief sequence \( \tilde{\pi}_T^g \). So equivalently we can calculate the smoothing distribution, \( p(\xi_t|\tilde{x}_T;\tilde{\pi}_T^g) \), using standard tools from Bayesian estimation. For nonlinear problems these algorithms will iterate a local approximation of the exact smoothing distribution \( p(\xi_t|\tilde{x}_T;\tilde{\pi}_T^g) \) in proximity of some reference trajectory \( \xi^i_T \). There exist several numerical strategies to execute this procedure amongst which the iEKS and iUKS. We refer to \cite{94} and \cite{95} and Appendix B for details and refer to the collection of possible algorithms as iterated RTS smoothers. The main steps are then summarised in Algorithm 3.

Again some remarks are in place. First we emphasise that if we would be able to calculate all probabilities exactly, there is no difference between this approach and

\(^\dagger\) The Differential Dynamic Programming or DDP algorithm is an extension of the iLQR algorithm. The algorithm is equivalent up to slightly more complex updates in the backward pass that take into account second order derivatives of the dynamics as opposed to the iLQR algorithm that only used the first order derivatives \cite{14,92,15,21}.
the one described in the previous paragraph. Nevertheless there will be differences once we use approximate inference methods such as the collection of iRTS smoothers proposed above. These iterated smoothers alternate between a filtering and smoothing pass which roughly correspond to the forward and backward passes described for the iELQR. However in the forward pass of the iELQR algorithm we only evaluate the current policy; we do not yet condition on the running posterior measurement sequence \( z^g \). The contrary is true for algorithm described here. This trajectory *innovation* concept (conditioning on the running posterior measurement sequence) is therefore unique to the present context and is equivalent to the Kalman update. The hypothesis is that now not only the policy and hence the control trajectory is optimised in the backward pass, but here also the state trajectory is attracted to more advantageous regions of the state space during the forward pass.

We emphasise that in order to implement the Kalman update (see Appendix B), an actual probabilistic emission model is required. We need a model in the likes of

\[
    z_t = h_t(\xi_t) + \zeta_t, \quad \zeta_t = \mathcal{N}(0, \Sigma_{\zeta\zeta,t}) \tag{61}
\]

Also note that by using an iterated RTS, we can only encode information about the cost rate function as well as the systems dynamics in the forward pass *but not* in the backward smoothing pass. Ideally our measurement model should therefore also incorporate first and second order gradient information about the cost function. E.g.

\[
p(z_t|\xi_t) \propto \exp(-r_t(\xi_t)) \\
\approx \exp \left( -\frac{1}{2} \left[ \begin{array}{c} 1 \\ \Delta \xi_t \end{array} \right]^T \left[ \begin{array}{cc} R_{\xi,\xi,t}^{-1} & 0 \\ 0 & \Sigma_{\zeta\zeta,t}^{-1} \end{array} \right] \left[ \begin{array}{c} 1 \\ \Delta \xi_t \end{array} \right] \right) \\
= \exp \left( -\frac{1}{2} \| \Delta \xi_t - R_{\xi,\xi,t}^{-1} R_{\xi,\zeta,t} \Sigma_{\zeta,\zeta,t}^{-1} \|_2^2 \right) \tag{62}
\]

Setting \( z_t = 0, h_t(\xi_t) \approx \Delta \xi_t - R_{\xi,\xi,t}^{-1} R_{\xi,\zeta,t} \Sigma_{\zeta,\zeta,t}^{-1} \) and \( \Sigma_{\zeta,\zeta,t}^{-1} = R_{\xi,\zeta,t} \) yields the required relations. Note that this model is not unique, any other decomposition would also do the trick.

### 6.2 Sample based trajectory optimisation

In addition to the previous paragraph and as a step-up to the present discussion, we note that sample based implementations of the RTS exist as well. Hence, nothing prevents us from applying such a method to obtain a sample based algorithm. We will not detail how such an algorithm would operate. However we grasp the opportunity to discuss a number of alternative sample based methods. In this paragraph we focus on algorithms that are closer related to reinforcement learning but that are still tailored to trajectory optimisation problems.

Again we will describe a number of procedures to update the linear Gaussian policy in \( \xi_t \). Though as opposed to the gradient based strategies described in the previous paragraph, here we desire to probe the problem locally not by relying on gradient information but by relying on stochasticity. This involves reinterpreting and executing the policy beliefs stochastically. Then a sufficiently rich set of \( M \) sample
paths, $D^g = \{\xi^m_T\}_{m=1}^M$, can be spawned by sampling from the policy beliefs $\pi$:

\[
x_0^m \sim p(x_0)
\]
\[
w_t^m \sim \mathcal{N}(u_t|k_t^m + K_t^mx_t^m, \Sigma_t^m)
\]
\[
x_{t+1}^m = f_t(x_t^m, u_t^m)
\]  

To put it differently, we execute the forward uncertainty propagation associated to the forward pass of the iELQR algorithm using a Monte Carlo approach. Using Monte Carlo approaches to probe the problem locally has some advantages over gradient based approaches. (1) A major benefit of using trajectory samples over gradient related approaches is that we do not need a dynamic system model or at least not a model that is differentiable. (2) Furthermore the computation of the different sample paths can be distributed. Then relying on the theoretical concepts from section 3, we can derive a number of strategies to perform an appropriate backward pass based and update the parametric policies. In that sense, the idea is closely related to particle filtering methods where a set of particles is used to represent the state belief.

6.2.1 Path Integral Control The first idea is directly inspired by recent work on PIC control methods [40]. Where historically this line of algorithms derived from the LSOC framework, the discovery of path integral expressions for the optimal policy in the setting of RSOC (see section 4.4) allows to generalise these methods overcoming restrictions that root back to the LSOC framework.

Technically the strategy boils down to a likelihood estimation procedure. At each time instant $t$ we infer parameters $\theta_t^g = \{k_t^g, K_t^g, \Sigma_t^g\}_t$ calculating the expected I-projection of a parametric policy $\tilde{\pi}_T(\xi_t; \theta_t)$ onto the theoretical policy $\pi_T(\xi_t)$. We define $p(x_t|\xi_0; \tilde{\pi}_T)$ as the state distribution at time $t$ under the given approximate policy belief sequence $\tilde{\pi}_T$. There is no theoretical motivation for this choice apart from that such data is available.

\[
\arg\min_{\theta_t} \mathbb{E}_{p(x_t|\xi_0; \tilde{\pi}_T)} \left[ \mathbb{D} \left[ \pi_{t+1}^g(\xi_t) \middle\| \pi_t(\xi_t; \theta_t) \right] \right]
\]

which is equivalent to

\[
\arg\max_{\theta_t} \mathbb{E}_{p(x_t|\xi_0; \tilde{\pi}_T)} \left[ \mathbb{E}_{p(u_t|x_t, \pi_{t+1}^g)} \left[ \log \pi_t(\xi_t; \theta_t) \right] \right]
\]  

Second we make use explicitly of the path integral expressions derived in section 4.4. Substitute the path integral expression for $\pi_{t+1}^g$ yields the following explicit update.

\[
\theta_t^{g+1} = \arg\max_{\theta_t} \mathbb{E}_{p(x_t|\xi_0; \tilde{\pi}_T)} \left[ e^{-r(\xi_t)} \log \pi_t(\xi_t; \theta_t) \right]
\]
\[
\approx \arg\max_{\theta_t} \sum_{m} \frac{w_t^m}{\pi_t^m} \log \pi_t(\xi_t|\pi_{t+1}^g)
\]  

† Superscripting a function with $m$ implies evaluation of the $m^{th}$ path, e.g. $f_t(x_t^m, u_t^m) = f_t^m$. For ease of exposition we limit our discussion to the hyper iterations, which would boil down to interrupting the main iterations after one update.
where $- \log w^m \propto R(\xi^m)$. The updates can be estimated accordingly. For conciseness we refer to Appendix Appendix C for details.

6.2.2 Backward pass through quadratic approximations Our second approach makes use of the observation that if the $Q$-function is quadratic, by definition the associated posterior policy is of the local Gaussian form (60). Hence the idea is here to extract a quadratic approximation of the $(g+1)^{th}$ $Q$-function from the sample set $D^g$ instead of using gradient information. Similar ideas were first described by [96, 97]. Ergo we require a second order approximation of the $Q$-function with $\phi = \{\hat{Q}_{\xi}, \hat{Q}_{\xi\xi}\}$.

$$Q(\xi; \phi) = \frac{1}{2} \begin{bmatrix} 1 \\ \xi \end{bmatrix}^T \begin{bmatrix} \hat{Q}_{\xi} \\ \hat{Q}_{\xi\xi} \end{bmatrix} \begin{bmatrix} 1 \\ \xi \end{bmatrix}$$  \hspace{1cm} (67)

To estimate parameters $\{\phi^g_t\}_t$ a second order least squares approach can be used. Thus now we require estimates for $Q^g_{t,m}$.

$$\phi^g_t = \arg \min_{\phi_t} \sum_m (Q(\xi^m_t; \phi_t) - Q^g_{t,m})^2$$  \hspace{1cm} (68)

Therefore we can make use of the recursive equations for the regularised value and $Q$-functions. Furthermore we will make explicit use of the fact that for deterministic dynamics the expression for the $Q$-function is related exactly to the expression of the $V$-function. As mentioned a quadratic surrogate of $Q^{g+1}_t$ can be obtained from samples $\{Q^{g+1}_t,m\}_m$. Then recall that

$$Q^{g+1}_{t,m} = x^m_t + V^{g+1}_{t+1,m}$$  \hspace{1cm} (69)

Thus the problem is reduces to finding an estimate of $V^{g+1}_{t+1,m}$. Given that we already have access to estimates $\hat{Q}^{g+1}_t$ and $Q^{g+1}_t$ at this point, we can exploit the quadratic expressions described in section 4.5 to evaluate $V^{g+1}_{t+1}$ and evaluate it at $x^m_{t+1}$. Therefore we can make use of the recursive equations for the regularised value and $Q$-functions. Furthermore we will make explicit use of the fact that for deterministic dynamics the expression for the $Q$-function is related exactly to the expression of the $V$-function. As mentioned a quadratic surrogate of $Q^{g+1}_t$ can be obtained from samples $\{Q^{g+1}_t,m\}_m$. Then recall that

$$Q^{g+1}_{t,m} = x^m_t + V^{g+1}_{t+1,m}$$  \hspace{1cm} (69)

Thus the problem is reduces to finding an estimate of $V^{g+1}_{t+1,m}$. Given that we already have access to estimates $\hat{Q}^{g+1}_t$ and $Q^{g+1}_t$ at this point, we can exploit the quadratic expressions described in section 4.5 to evaluate $V^{g+1}_{t+1}$ and evaluate it at $x^m_{t+1}$. It follows that this procedure, in combination with the LQER expressions for the local Gaussian policy, implements a sample based approximation of the iELQR method described in the previous paragraph. The benefit is that here we do not limit the expressiveness of the forward uncertainty propagation to some approximate mechanism such as the EKF of UKF. Instead we make use of a full Monte Carlo method. Then because we rely on the linear-quadratic updates from the ELQR to determine the policy and to evaluate the value function, we indirectly compute ensemble gradient approximations of the local dynamics and cost model.

7 Conclusion

In this paper we have attempted to lay bare a probabilistic inference mechanism underlying existing inference principles in optimal control, therewith providing a deontological argument for the use of information-theoretic measures in this context as opposed to the teleological arguments that are commonly pleaded. Our investigation

\*\* An alternative strategy would involve parameterising the entire path probability distribution, $p(\Xi_T|x_0; \pi^{(g)}_T)$, by substituting a parameterised policy series $\{\pi_t(u_t|x_t; \theta_t)\}$ and use the I-projection strategy on the entire posterior path probability $p(\Xi_T|x_0; \pi^{(g+1)}_T)$. It is easily verified that this strategy ultimately results into the strategy above except for the likelihood weights $w^m_t$ that are now time invariant and thus equivalent for every temporal policy update.
has lead to the conclusion that the Stochastic and Risk Sensitive Optimal Control problem can be addressed making use of the Expectation-Maximisation algorithm. The governing update rules are shown to constitute a consistent inference mechanism to recover the deterministic policy in a finite amount of iterations. In other words our analysis points out that by iterating the derived updates we may infer the actual deterministic optimal policy that renders extreme the two underlying optimal control problems.

It is our contention that these observations complement existing Control as Inference approaches, as has been discussed by many authors [8, 6, 35, 40, 38, 91, 39], with a deontological argument of validity. Therefore we argue that the exposed mechanism can be interpreted as an extension of the traditional optimal control paradigm and identify it as an instance of probabilistic numerics, in the sense that policy belief functions are generated that express our uncertainty about an underlying deterministic solution. Therefore we propose to refer to the ensemble paradigm as entropy regularised optimal control.

What is also remarkable is that the entropy regularised reciprocal of the risk-sensitive optimal control problem allows to express the iterate policy as a path integral in the sense of Feynman and Kac. This is a characteristic that was reserved to the Linearly Solvable Optimal Control framework until now or entropy regularised trajectory optimisation problems. However the latter problem formulations are lifted entirely to the state-space [35], limiting their application potential. Furthermore this observation solves the question raised by [39] who sought a connection between Path Integral Control and message passing based on the problem that we would now refer to as the entropy regularised reciprocal of the risk sensitive optimal control problem.

Based on our theoretical investigation an outlook was given for application of the paradigm in trajectory optimisation. We hope this outlook may inspire a new family of algorithms with improved numerical characteristics that benefit from the insights provided by this work.

In conclusion we note that our investigation has been limited to optimal control problems in discrete time. Attempting a generalisation to continuous-time problems seems worthwhile both from a theoretical as well as a practical point of view. However it appears challenging to generalise the use of the Expectation-Maximisation algorithm to continuous time optimal control problems. Though future work may draw inspiration from the rich literature on stochastic calculus.

Appendix A Recursive solution of SOC & RSOC

In this appendix we give some additional details on the standard stochastic and risk sensitive optimal control problems that were introduced in section 2.1.

Appendix A.1 Backward recursive expressions

It can be shown that the solution of the SOC and RSOC problems are governed by the following equations. Here $\pi_t$ denotes the deterministic policy and $V_t(x_t)$ denotes the value function.

$$\pi_t(x_t) = \arg \min_{u_t} Q_t(\xi_t)$$

$$V_t(x_t) = \min_{u_t} Q_t(\xi_t)$$

(A.1)
On Optimal Control and Expectation-Maximisation

Clearly either function is governed by the $Q$-function. The definition of the $Q$-function determines whether we are dealing with the SOC or RSOC setting.

**SOC**

\[
Q_t(\xi_t) = r_t(\xi_t) + \mathbb{E}_{p(x'_{t+1}|\xi_t)}[V_{t+1}(x')]
\]

(A.2)

**RSOC**

\[
Q_t(\xi_t) = r_t(\xi_t) - \log \mathbb{E}_{p(x'_{t+1}|\xi_t)}[\exp(-V_{t+1}(x'))]
\]

(A.3)

For completeness and ease of comparison we can summarise these results in a similar table as was provided in Table 1.

### Appendix A.2 LQR & LQER

In this paragraph, let us specialise the results above to the Linear-Quadratic-Regulator (LQR) and Linear-Exponential-Quadratic-Regulator (LEQR) problem. Again these are examples of the SOC and RSOC problem specialised to linear-Gaussian dynamics, $p(x_{t+1}|\xi_t) = \mathcal{N}(x_{t+1}|F_{\xi,t}\xi_t + f_t, P_t)$, and Quadratic costs, $r_t(\xi_t) = \frac{1}{2}\xi_t^T R_{\xi,t}\xi_t + \xi_t^T R_{x,t} + \ldots$ and $r_T(x_T) = \frac{1}{2}\xi_T^T R_{x,x,T}x_T + x_T^T R_{x,T} + \ldots$. Here also the policy is linear affine in the state and both the $V$- and $Q$-function are quadratic in their arguments.

Using the same notation as before one easily verifies the following solution

**policy $\pi$**

\[
k_t = -Q^{-1}_{uu,t} Q_{u,t}
\]

\[
K_t = -Q^{-1}_{uu,t} Q_{ux,t}
\]

**$V$-function**

\[
V_{x,t} = Q_{x,t} + Q_{u,t}k_t + K_t^T Q_{u,u,t}k_t
\]

\[
V_{xx,t} = Q_{xx,t} + K_t^T Q_{u,x,t} + Q_{x,u,t}K_t + K_t^T Q_{u,u,t}K_t
\]

(A.5)

As noted the policy and value function are equivalent irrespective of the OC framework. The $Q$-function determines whether we are dealing with the one or the other.

**LQR**

\[
Q_{\xi,t} = R_{\xi,t} + F_{\xi,t}^T (V_{xx,t+1}f_t + V_{x,t})
\]

\[
Q_{\xi\xi,t} = R_{\xi\xi,t} + F_{\xi,t}^T V_{xx,t+1}F_{\xi,t}
\]

(A.6)

**LQER**

\[
Q_{\xi,t} = R_{\xi,t} + F_{\xi,t}^T (V_{xx,t+1}^{-1} + P_t)^{-1} (V_{xx,t+1}^{-1}V_{x,t} + f_t)
\]

\[
Q_{\xi\xi,t} = R_{\xi\xi,t} + F_{\xi,t}^T (V_{xx,t+1}^{-1} + P_t)^{-1} F_{\xi,t}
\]

(A.7)
Appendix B  Bayesian Filtering & Smoothing of CHMMs

We consider a Controlled Hidden Markov Model (CHMM) which is an extension
of the standard HMM where we a hidden action sequence, $u_T$. Recall figure 2.

To further specify the probabilistic state-space model we characterise the following
set of conditional probability distributions describing the probabilistic dynamics
of the system. The model is Markovian, mathematically this property manifests
as conditional independence on historical measurements (not necessarily future
measurements). We have that $p(x_t|\xi_{t-1}, z_{t-1}) \equiv p(x_t|\xi_{t-1})$, $p(u_t|x_t, z_{t-1}) \equiv p(u_t|x_t)$
and $p(z_t|\xi_t, z) \equiv p(z_t|\xi_t)$. As a result, when the system manifests into a path \{\xi, z\},
it does so according to the probabilities determined by the model $M$.

\[
M = \begin{cases} 
  p(x_0) & \text{(initial model)} \\
  p(x_t|\xi_{t-1}) & \text{(transition model)} \\
  p(u_t|x_t) & \text{(control model)} \\
  p(z_t|\xi_t) & \text{(emission model)} 
\end{cases}
\]  

(B.1)

Two key problems in Bayesian estimation are then determined by the calculation
of the marginal joint probabilities conditioned on the leading, $z_t$, and full, $z_T$,
measurement sequences. These conditional probabilities are better known as the
filtering, $p(\xi_t|z_t)$, and smoothing, $p(\xi_t|z_T)$, densities respectively. The exposition in
most reference works is limited to regular HMMs, excluding the control variable and
probabilistic policy. In the following, therefore we treat filtering and smoothing in the
more general setting of controlled Hidden Markov Models.

Appendix B.1  Filtering

The purpose of Bayesian filtering is to compute the joint marginal posterior (or so-
called filtering density) of the joint state-action, $\xi_t$, at each time step, $t$, given the
history of the measurements up to the time step $t$, i.e. $p(\xi_t|z_t)$.

Appendix B.1.1  Exact filtering  There exists a well known and exact forward
recursion to compute the filtering distributions. The general procedure is described
next for controlled HMMs. An important intermediate result of the computation is
the so-called predictive joint distribution which is defined as $p(\xi_t|z_{t-1})$. By definition
we have that $p(\xi_0|z_{-1}) = p(\xi_0)$.

initialisation  The recursion is initialised with $p(\xi_0)$.

\[
p(\xi_0) = p(u_0|x_0)p(x_0)
\]  

(B.2)

update step  Given a measurement $z_t$ and the predictive joint probability $p(\xi_t|z_{t-1})$,
the present marginal joint posterior, $p(\xi_t|z_t)$, can be computed using Bayes' rule.

\[
p(\xi_t|z_t) \propto p(z_t|\xi_t)p(\xi_t|z_{t-1})
\]  

(B.3)

forward prediction step  Finally a two step computation can be used to determine
the next predictive joint probability provided with the current marginal joint
posterior. First the predictive state probability $p(x_{t+1}|z_t)$ is determined. Second
the predictive joint probability can be computed.
\[ p(x_{t+1}|\hat{z}_t) = \int p(x_{t+1}|\xi_t)p(\xi_t|\hat{z}_t)d\xi_t \]

\[ p(\xi_{t+1}|\hat{z}_t) = p(u_{t+1}|x_{t+1})p(x_{t+1}|\hat{z}_t) \]  

(B.4)

It follows that the general filtering densities can be determined in a single forward recursive calculation. In a linear-Gaussian this procedure can be evaluated explicitly. This result is known as the Kalman filter. For brevity we omit explicit equations.

**Appendix B.1.2 Approximate filtering** It often happens in practical applications that the dynamic and measurement models are not linear so that the Kalman filter does not apply. However, often the filtering densities of this type of model can still be approximated by Gaussian densities. Within the literature there are two main strategies to approximate such Gaussian densities. Either strategy roots back to the practice of a different uncertainty quantification strategy to locally approximate the forward probabilities \(p(x_t|\xi_{t-1}), p(u_t|x_t)\) and \(p(z_t|\xi_t)\). Extended Kalman Filters (EKF) practice a Taylor approximation to linearise the the forward probabilities. Unscented Kalman Filters practice the unscented transform to propagate the uncertainty. Essentially these algorithms compute an (n) (ensemble) gradient based linear approximation about some reference trajectory, similar to how the LQR algorithm can be used to solve generic nonlinear OCPs. Interestingly here the reference trajectory is determined on the fly, coinciding with the expected value of the joint marginal posterior. Hence there is not iterative aspect to these procedures except for the update step, where in some implementations of the EKF or UKF this procedure is iterated to obtain a better approximation of the marginal joint posterior. In that sense these iterations are local and do not reconsider the entire trajectory, only updating the local estimate of \(p(\xi_t|\hat{z}_t)\). See [94, 95] for details on approximate filtering.

**Appendix B.2 Smoothing**

The purpose of Bayesian smoothing is to compute the marginal joint posterior (or so-called smoothing densities) of the state-action, \(\xi_t\), at the time step, \(t\), after receiving the measurements up to a time step \(T\) where \(T > t\), i.e. \(p(\xi_t|z_T)\).

**Appendix B.2.1 Exact smoothing** There exists a well known and exact forward-backward recursive procedure to compute the smoothing densities. The general procedure is described next for controlled HMMs. An important intermediate result of the computation are the filtering densities, which determines the forward recursion, and the associated predicted state probabilities \(p(x_{t+1}|\hat{z}_t)\). In a backward recursion the filtering density \(p(\xi_t|\hat{z}_t)\) is then corrected additionally conditioning onto the future measurements \(z_{t+1}\) not considered in filtering. By definition we have that \(p(\xi_T|z_T)\) qualifies both as a filtering and smoothing density.

**initialisation** The recursion is initialised with the filtered density \(p(\xi_T|z_T)\).

**backward correction step** Next the present future measurements are encoded into the present filtering density, \(p(\xi_t|\hat{z}_t)\), computing the present smoothing density, \(p(\xi_t|z_T)\). Note that the correction step makes use of the trailing smoothing marginalised distribution \(p(x_{t+1}|z_T)\) and the predicted state density, \(p(x_{t+1}|\hat{z}_t)\).
\[ p(\xi_t|z_T) = p(\xi_t|z_t) \int \frac{p(x_{t+1}|z_T)}{p(x_{t+1}|z_t)} p(x_{t+1}|\xi_t) dx_{t+1} \]  

(B.5)

It follows that the general smoothing densities can be determined in a single forward recursive calculation, determining the filtering distributions, and, a single backward recursive calculation. In a linear-Gaussian setting this procedure can be evaluated explicitly. This result is known as the Rauch-Tung-Striebel (RTS) smoother.

**Appendix B.2.2 Approximate smoothing** Similar to the filtering problem, it often happens in practical applications that the dynamic and measurement models are not linear so that the RTS smoother simply does not apply. Often also the smoothing densities can be approximated with a Gaussian density. The same uncertainty quantifications procedure can be used as described for the filtering problem which establishes an approximate calculation for the forward recursion. Provided with the local linearisation of the problem, the linear backward smoothing procedure can now readily be applied. This idea characterised the standard Extended (EKS) and Unscented Kalman Smoother (UKS). Interestingly here an iterative procedure can be established since in general the smoothed trajectory will not coincide with the filtered trajectory so that the linearisation point can be reconsidered after every update. In that sense the linear forward and backward recursive procedure can be alternated to improve the local approximation. This idea characterised the iterated EKF (iEKS) and iKF (iUKS). See [94, 95].

**Appendix C Exact Path Integral Control update**

For completeness we give here the explicit solution of the optimisation problem in (66). The parameters are given by

\[
\begin{align*}
  k^{g+1}_t &= \hat{\mu}^{g+1}_x - K^{g+1}_{x,t} \hat{\mu}^{g+1}_x \\
  K^{g+1}_t &= \Sigma^{g+1}_{u,t} - K^{g+1}_{x,t} \Sigma^{g+1}_{xu,t} K^{g+1}_{x,t} \\
  \Sigma^{g+1}_{\xi,t} &= \Sigma^{g+1}_{u,t} - K^{g+1}_{x,t} \Sigma^{g+1}_{xu,t} K^{g+1}_{x,t} \Sigma^{g+1}_{x,x,t} K^{g+1}_{x,t} \\
\end{align*}
\]

(C.1)

where \( \langle (\cdot) \rangle = \sum_k \frac{w^k_t}{\sum_k w^k_t} (\cdot) \) and

\[
\begin{align*}
  \hat{\mu}^{g+1}_{\xi,t} &= \langle \xi_t \rangle \\
  \hat{\Sigma}^{g+1}_{\xi,t} &= \left( \xi_t - \hat{\mu}^{g+1}_{\xi,t} \right) \left( \xi_t - \hat{\mu}^{g+1}_{\xi,t} \right)^T \\
\end{align*}
\]

(C.2)

Based on the structural properties of the updates we can conclude that this strategy uses a likelihood weighted strategy to estimate a Gaussian approximation of the marginal distributions \( p(\xi_t|x_0; \tilde{z}_T^{g+1}) \) and then conditions this estimate on \( x_t \) to obtain an estimate of the updated policy, \( \tilde{\pi}_T^{g+1} \). This is actually similar to the procedure proposed by the iELQR. In addition we note that this approach extends to stochastic system dynamics, in which case we have to adopt a stochastic simulation of the system according to \( x^k_{t+1} \sim p(x_{t+1}|\xi^k_t) \). As discussed in section 4.4 then this approach works only for the risk sensitive setting.
On Optimal Control and Expectation-Maximisation

References

[1] Mordatch I, Todorov E and Popović Z 2012 ACM Transactions on Graphics 31 ISSN 0730-0301 URL https://doi.org/10.1145/2185520.2185539
[2] Toussaint M 2015 Logic-geometric programming: An optimization-based approach to combined task and motion planning Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI)
[3] Posa M, Cantu C and Tedrake R 2014 The International Journal of Robotics Research 33 69–84 (Preprint) URL https://doi.org/10.1177/0278364913506757
[4] Oates C and Sullivan T 2019 Statistics and computing 29 1335–1351 ISSN 0960-3174
[5] Hennig P, Osborne M and Girolami M 2015 Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 471 20150142
[6] Rawlik K, Toussaint M and Vijayakumar S 2013 On stochastic optimal control and reinforcement learning by approximate inference Twenty-Third International Joint Conference on Artificial Intelligence
[7] Levine S and Koltun V 2013 Variational policy search via trajectory optimization Advances in Neural Information Processing Systems vol 26 ed Burges C J C, Bottou L, Welling M, Ghahramani Z and Weinberger K Q (Curran Associates, Inc.) URL https://proceedings.neurips.cc/paper/2013/file/38af86134b65d0f10fe33d30dd76442e-Paper.pdf
[8] Levine S 2018 arXiv preprint arXiv:1805.00909
[9] Jaynes E T 2003 Probability Theory: The Logic of Science (Cambridge University Press)
[10] Bishop M 2006 Pattern Recognition and Machine Learning (Information Science and Statistics) (Berlin, Heidelberg: Springer-Verlag) ISBN 0387310738
[11] Caticha A 2011 Entropic inference AIP Conference Proceedings vol 1305 (American Institute of Physics) pp 20–29
[12] Caticha A 2013 Entropic inference: Some pitfalls and paradoxes we can avoid AIP Conference Proceedings vol 1553 (American Institute of Physics) pp 200–211
[13] Thomas M and Joy A T 2006 Elements of information theory (Wiley-Interscience)
[14] Kullback S 1959 Information theory and statistics (John Wiley)
[30] Jaynes E and Justice J H 1986 Bayesian Methods: General Background (Cambridge University Press) p 1–25
[31] Jaynes E 1982 Proceedings of the IEEE 70 939–952
[32] Murphy K P 2012 Machine learning: a probabilistic perspective (MIT press)
[33] Nielsen F 2018 Notices of the AMS 65 321–324
[34] Luo X 2019 arXiv preprint arXiv:1812.03457
[35] Lefebvre T and Crevecoeur G 2020 Entropy 22 1120
[36] Peters J, Mulling K and Altun Y 2010 Relative entropy policy search Twenty-Fourth AAAI Conference on Artificial Intelligence
[37] Schulman J, Levine S, Abbeel P, Jordan M and Moritz P 2015 Trust region policy optimization International Conference on Machine Learning pp 1889–1897
[38] Toussaint M 2009 Robot trajectory optimization using approximate inference Proceedings of the 26th annual international conference on machine learning pp 1049–1056
[39] Watson J 2021 Reinforcement Learning Algorithms: Analysis and Applications 189–206
[40] Lefebvre T and Crevecoeur G 2021 Entropy regularized deterministic optimal control: Path integral solution and control method (Preprint [1xx.yyyyy])
[41] Todorov E 2007 Linearly-solvable markov decision problems Advances in neural information processing systems pp 1369–1376
[42] Todorov E 2009 Proceedings of the national academy of sciences 106 11478–11483
[43] Djuric J and Todorov E 2012 Linearly solvable optimal control Reinforcement learning and approximate dynamic programming for feedback control vol 17 (Citeseer) pp 119–141
[44] Lefebvre T and Crevecoeur G 2019 Path integral policy improvement with differential dynamic programming 2019 IEEE International Conference on Advanced Intelligent Mechatronics (AIM) (IEEE)
[45] Abdolmaleki A, Lioutikov R, Peters J, Lau N, Reis L and Neumann G 2015 Model-based relative entropy stochastic search Advances in Neural Information Processing Systems pp 3537–3545
[46] Abdolmaleki A, Price B, Lau N, Reis L and Neumann G 2017 Deriving and improving cmac models with information geometric trust regions Proceedings of the Genetic and Evolutionary Computation Conference (ACM) pp 657–664
[47] Kalman R 1960 Journal of Basic Engineering 82 35–45 ISSN 0021-9223 (Preprint https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf) URL https://doi.org/10.1115/1.3662552
[48] Ho Y and Lee R 1964 IEEE transactions on automatic control 9 333–339
[49] Pearson J 1966 SIAM Journal on Control 4 594–600
[50] Pavon M and Wets R 1982 The duality between estimation and control from a variational viewpoint: The discrete time case Algorithms and Theory in Filtering and Control (Springer) pp 1–11
[51] Todorov E 2008 General duality between optimal control and estimation Decision and Control, 2008. CDC 2008. 47th IEEE Conference on (IEEE) pp 4286–4292
[52] Kappen H 2005 Physical review letters 95 200201
[53] Kac M 1951 On some connections between probability theory and differential and integral equations Proceedings of the second Berkeley symposium on mathematical statistics and probability (The Regents of the University of California)
[54] Feynman R P 2005 Space-time approach to non-relativistic quantum mechanics Feynman’s Thesis—A New Approach To Quantum Theory (World Scientific) pp 71–109
[55] Kappen H, Wiegerinck W and van den Broek B 2007 Autonomous Agents and Multi-Agent Systems
[56] Todorov E 2009 Eigenfunction approximation methods for linearly-solvable optimal control problems Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL’09. IEEE Symposium on (IEEE) pp 161–168
[57] Toussaint M and Storkey A 2006 Probabilistic inference for solving discrete and continuous state markov decision processes Proceedings of the 23rd international conference on Machine learning pp 945–952
[58] Toussaint M 2009 Pros and cons of truncated gaussian ep in the context of approximate inference control NIPS-workshop on probabilistic approaches for robotics and control vol 21
[59] Peters J 2015 RI Seminar: Jan Peters : Motor Skill Learning: From Simple Skills to Table Tennis and Manipulation URL https://www.youtube.com/watch?v=73s2dzAr5G_4&t=1373s
[60] Kakade S 2001 Advances in neural information processing systems 14
[61] Bagnell J and Schneider J
[62] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O 2017 arXiv preprint arXiv:1707.06347
On Optimal Control and Expectation-Maximisation

[63] Ziebart B, Maas A, Bagnell J and Dey A 2008 Maximum entropy inverse reinforcement learning. Aaai vol 8 (Chicago, IL, USA) pp 1433–1438

[64] Levine S and Koltun V 2013 Guided policy search International conference on machine learning (PMLR) pp 1–9

[65] Haarnoja T, Tang H, Abbeel P and Levine S 2017 Reinforcement learning with deep energy-based policies Proceedings of the 34th International Conference on Machine Learning-Volume 70 (JMLR. org) pp 1352–1361

[66] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P et al. 2018 arXiv preprint arXiv:1812.05905

[67] Theodorou E, Buchli J and Schaal S 2010 Journal of Machine Learning Research 11 3137–3181

[68] Theodorou E, Buchli J and Schaal S 2010 Reinforcement learning of motor skills in high dimensions: A path integral approach 2010 IEEE International Conference on Robotics and Automation (IEEE) pp 2397–2403

[69] Stulp F and Sigaud O 2012 Path integral policy improvement with covariance matrix adaptation Proceedings of the 29th International Conference on Machine Learning (ICML-12)

[70] Theodorou E, Krishnamurthy D and Todorov E 2013 From information theoretic dualities to path integral and kullback-leibler control: Continuous and discrete time formulations The Sixteenth Yale Workshop on Adaptive and Learning Systems

[71] Theodorou E 2015 Entropy 17 3352–3375

[72] Gómez V, Kappen H, Peters J and Neumann G 2014 Policy search for path integral control Joint European conference on machine learning and knowledge discovery in databases (Springer) pp 482–497

[73] Pan Y, Theodorou E and Kontitsis M 2015 Sample efficient path integral control under uncertainty Advances in neural information processing systems pp 2314–2322

[74] Kappen H and Baio H C 2016 Journal of statistical physics 162 1244–1266

[75] Thalmeier D, Kappen H, Totaro S and Gómez V 2020 arXiv preprint arXiv:2005.06364

[76] Williams G, Aldrich A and Theodorou E 2017 Journal of Guidance, Control, and Dynamics 40 344–357

[77] Williams G, Drews P, Goldfain B, Rehg J and Theodorou E 2018 IEEE transactions on robotics 34 1603–1622

[78] Rajamäki J, Naderi K, Kyrki V and Hämäläinen P 2016 Sampled differential dynamic programming 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE) pp 1402–1409

[79] Colin S, Kendall L, Aravind R, Siddhartha S and Emanuel T 2020 Lyceum: An efficient and scalable ecosystem for robot learning (Preprint 2001.07343)

[80] Nagabandi A, Konolige K, Levine S and Kumar V 2020 Deep dynamics models for learning dexterous manipulation Conference on Robot Learning (PMLR) pp 1101–1112

[81] Williams G, Drews P, Goldfain B, Rehg J and Theodorou E 2016 Aggressive driving with model predictive path integral control 2016 IEEE International Conference on Robotics and Automation (ICRA) pp 1433–1440

[82] Kahn G, Abbeel P and Levine S 2021 IEEE Robotics and Automation Letters 6 1872–1879

[83] Kahn G, Abbeel P and Levine S 2021 IEEE Robotics and Automation Letters 6 1312–1319

[84] Chebotar Y, Kalakrishnan M, Yahya A, Li A, Schaal S, Haarnoja T, Byravan A, Nair V, Neumann G et al. 2017 Path integral guided policy search 2017 IEEE international conference on robotics and automation (ICRA) (IEEE) pp 3381–3388

[85] Neumann G et al. 2011 Variational inference for policy search in changing situations Proceedings of the 28th international conference on machine learning, ICML 2011 pp 817–824

[86] Abdolmaleki A, Springenberg J, Tassa Y, Munos R, Heess N and Riedmiller M 2018 Maximum a posteriori policy optimisation International Conference on Learning Representations URL https://openreview.net/forum?id=S1ANxQ0W6

[87] Lee A X, Devin C M, Zhou Y, Lampe T, Boumal K S, Springenberg J T, Byravan A, Abdolmaleki A, Gileadi N, Kosid D, Fantacci C, Chen J E, Raju A, Jeong R, Neunert M, Lauren E, Saliceti S, Casarini F, Riedmiller M, raia hadsell and Nori F 2021 Beyond pick-and-place: Tackling robotic stacking of diverse shapes 5th Annual Conference on Robot Learning URL https://openreview.net/forum?id=U0Q8CrtBJxJ

[88] Liu S, LeVer G, Wang Z, Merel J, Eslami S M A, Hennes D, Czarnecki W M, Tassa Y, Omidshafiei S, Abdolmaleki A, Siegel N Y, Hasenclever L, Marris L, Tenyasuvunakool S, Song H F, Wulfmeier M, Muller P, Haarnoja T, Tracey B D, Tuyls K, Graepel T and Heess N 2021 From motor control to team play in simulated humanoid football (Preprint 2105.12196)

[89] Imohiosen A, Watson J and Peters J 2020 Active inference or control as inference? a unifying view International Workshop on Active Inference (Springer) pp 12–19
[90] Watson J and Peters J 2021 Advancing trajectory optimization with approximate inference: Exploration, covariance control and adaptive risk American Control Conference (ACC)
[91] Watson J, Abdulsamad H, Findeisen R and Peters J 2021 Stochastic control through approximate bayesian input inference (Preprint 2105.07693)
[92] Tassa Y, Erez T and Todorov E 2012 Synthesis and stabilization of complex behaviors through online trajectory optimization 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE) pp 4906–4913
[93] Manchester Z and Kuindersma S 2016 Derivative-free trajectory optimization with unscented dynamic programming 2016 IEEE 55th Conference on Decision and Control (CDC) (IEEE) pp 3642–3647
[94] Särkkä S 2013 Bayesian filtering and smoothing 3 (Cambridge University Press)
[95] Barfoot T 2017 State estimation for robotics (Cambridge University Press)
[96] Akrour R, Neumann G, Abdulsamad H and Abdolmaleki A 2016 Model-free trajectory optimization for reinforcement learning International Conference on Machine Learning (PMLR) pp 2961–2970
[97] Akrour R, Abdolmaleki A, Abdulsamad H, Peters J and Neumann G 2018 The Journal of Machine Learning Research 19 565–589