2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Adaptive Risk-Sensitive Optimal Control with Dual Features through Active Inference

Mohammad Mahmoudi Filabadi, Tom Lefebvre and Guillaume Crevecoeur

***Abstract*—We propose a tractable adaptive risk-sensitive optimal control framework tailored to uncertain nonlinear stochastic system dynamics. The architecture exhibits dual features meaning that the controller actively maintains a balance between exploitation and exploration. The problem statement is cast as an instance of Active Inference. Active Inference is an emerging framework in theoretical neuroscience that seeks to explain the behaviour of biological agents by practising inference on probabilistic graph models. The developed algorithm leverages a receding horizon strategy that simultaneously estimates the uncertain parameters of the dynamic system from past observations and designs controller parameters by predicting the future performance of the controlled system. The algorithm does not make use of the separation and certainty equivalence principles. We further show that for the special case of linearly parameterized controller and dynamics, the approach leads to a quadratic programming problem maintaining a manageable computational complexity. The capability and anticipated properties of the proposed algorithm are demonstrated on a simulated nonlinear system.**

I. Introduction

Risk-sensitive optimal control (RSOC) refers to a generalization of conventional stochastic optimal control theory by incorporating some notion of risk, e.g. higher moments of the conventional utility function, into the performance measure [1], [2]. Although the concept has been known for several decades and is well-studied theoretically [3], only recently a small number of contributions have been investigating practical implementations for non-trivial cases that reach beyond the linear-quadratic setting. E.g. in [2], an iterative nonlinear RSOC with imperfect observations was proposed using iterative linearization of nonlinear system dynamics and employing the linear exponential quadratic Gaussian (LEQG) controller [3]. Other works have explored how to deal with uncertain dynamics and adaptivity. An adaptive risk-sensitive model predictive control with stochastic search is suggested by [4]. In [5], [6], the RSOC problem is solved using value iteration and Q-learning. Here no model is required at all.

In terms of balancing between exploration and exploitation, adaptive controllers can be classified into two categories - adaptive controllers with or without dual features. Dual control theory, founded by the seminal work of Feldbaum [7], [8], focuses on the idea that controlling an unknown system imposes a dual objective. The primary goal is to control the system effectively based on some performance metrics while the controller should also intentionally explore the system to gather information in the short term for long-term performance improvement. This approach is referred to as *goal-directed* exploration, where the control system actively incorporates the learning process into its decision-making. In contrast, with non-dual adaptive controllers, learning is *accidental* or *passive*, resulting in unintentional exploration. Unfortunately, the optimal adaptive dual control is analytically intractable and methods need to be devised that solve an approximation of the original problem.

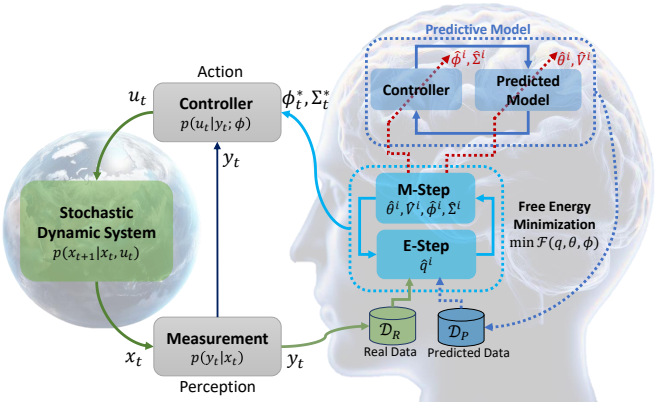

Fig. 1: The schematic diagram of the proposed algorithm.

This paper deals with the challenge of solving the adaptive RSOC problem with dual features by employing a probabilistic inference perspective. Expressing the stochastic optimal control problem (SOC) as a probabilistic inference problem has drawn significant attention from researchers [1], [9]. For instance, different approaches based on input inference for control (I2C) algorithm are developed in [10]–[12], which formulate the SOC problem as an input estimation problem and solve it using an expectation-minimization (EM) algorithm. Approximate inference control (AICO) [13], [14] is another approach to tackle the SOC problem. It interprets the quadratic cost over the control input as a prior distribution and computes the posterior distribution of states using the Gaussian message-passing technique.

In this paper, we address the aforementioned research gaps by extending the I2C algorithm [10]–[12] by incorporating the system identification which produces a tractable optimal adaptive dual controller. This is achieved by casting the problem in the active inference framework (AIF). AIF is a theoretical framework that explains the human brain's behaviour, including perception, planning, and action, in terms of probabilistic inference [15], [16]. In our approach, we predict the future behaviour of the controlled system using the collected information at each time step in a receding horizon strategy and apply the EM algorithm to minimize variational

M. M. Filabadi, T. Lefebvre and G. Crevecoeur are with the Dynamic Design Lab (D2Lab) of the Department of Electromechanical, Systems and Metal Engineering, Ghent University, B-9052 Ghent, Belgium e-mail: {mohammad.mahmoudifilabadi, tom.lefebvre, guillaume.crevecoeur}@ugent.be.
M. M. Filabadi, T. Lefebvre and G. Crevecoeur are member of core lab MIRO, Flanders Make, Belgium.

free energy bound by updating the unknown parameters and computing the smoothed density as a posterior distribution of states and control inputs over time (see Fig. 1). We will show that this concurrency between controller design and system identification intrinsically keeps the balance between exploitation and exploration. Furthermore, our approach only requires solving a quadratic program at every time step which enjoys a low computational cost.

The main contributions of the paper are summarized as

- Casting the adaptive RSOC problem as an active inference framework.
- Employing receding horizon strategy alongside active inference framework to implement the proposed algorithm online.
- Keeping exploitation and exploration balance effectively to achieve dual features.

We organize the remainder of the paper as follows. In Section II, we explain the problem formulation for designing the adaptive RSOC problem, and Section III represents its corresponding probabilistic inference problem. The proposed algorithm based on active inference is explained in Section IV. A simulation example is provided in Section V. In the end, Section VI is devoted to conclusions and future work.

II. Problem Statement

In this section, we introduce the dual control problem that will enjoy our attention throughout. As explained in the introduction the dual control problem falls apart into two subproblems that need to be solved simultaneously so that the overall control architecture exhibits dual features. The first subproblem is concerned with model-based control design. The second subproblem is concerned with identifying the parametric uncertainty of the dynamic system.

Consider a class of dynamical uncertain nonlinear stochastic discrete-time systems described by

$$x_{t+1} = f\left(x_t, u_t; \theta\right) + v_t \tag{1}$$

$$y_t = x_t + w_t \tag{2}$$

Here $x_t \in \mathbb{R}^{n_x}$, $y_t \in \mathbb{R}^{n_x}$, and $u_t \in \mathbb{R}^{n_u}$ represent the state, measurement output and control input vector of the dynamic system at discretized time t, $\theta \in \mathbb{R}^{n_\theta}$ represents a lumped parameter vector, $v_t \in \mathbb{R}^{n_x}$ and $w_t \in \mathbb{R}^{n_x}$ denote additive zero-mean Gaussian noises, which model the stochasticity of the system and measurement noise, respectively, given by Gaussian distributions $\mathcal{N}\left(v_t|0, V\right)$ and $\mathcal{N}\left(w_t|0, W\right)$ in which $V \in \mathbb{R}^{n_x \times n_x}$ and $W \in \mathbb{R}^{n_x \times n_x}$ are positive-definite covariance matrices, and $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_x}$ is a Lipschitz nonlinear mapping and its differentiability at θ will be required. The initial state is assumed to follow a normal distribution $x_0 \sim \mathcal{N}\left(x_0|m_{x_0}, P_{x_0}\right)$. Hereafter, we use $\tau_t \triangleq (x_t, u_t) \in \mathbb{R}^{n_x+n_u}$ to denote the concatenated state-control vector at time t.

A. Control subproblem

Remark that we assume the full state is observed; however, the state measurement itself is disturbed by additive noise. In the strictest interpretation, the system is therefore partially observed and we can only estimate the state. However, since the full state *is* measured, in this work, we aim to design a parameterized stochastic controller that is a function of the present measurement, y_t, alone. Specifically, we consider a class of parameterized controllers as follows

$$u_t = \pi(y_t; \phi) + \varepsilon_t \tag{3}$$

in which $\phi \in \mathbb{R}^{n_\phi}$ is the parameter vector of the controller that needs to be designed and $\varepsilon_t \in \mathbb{R}^{n_u}$ denotes an additive zero-mean Gaussian noise. The additive noise models the intentional stochasticity of the controller, its nature and purpose will become apparent later. The noise is given by a Gaussian distribution $\mathcal{N}\left(\varepsilon_t|0, \Sigma\right)$ in which $\Sigma \in \mathbb{R}^{n_u \times n_u}$ is a positive-definite covariance matrix which is absorbed in ϕ for design. Finally, $\pi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\phi} \mapsto \mathbb{R}^{n_u}$ represents a Lipschitz nonlinear mapping, differentiable in ϕ .

To design the controller, the following RSOC problem [1], [3] will be taken into account

$$\min_{\phi} \underbrace{-\frac{1}{\gamma}\log\left(\mathbb{E}_{p(\tau_{0:T}, y_{0:T}|\theta,\phi)}\left[\exp\left(-\gamma L\left(\tau_{0:T}\right)\right)\right]\right)}_{\triangleq J_\gamma(\theta,\phi)} \tag{4}$$

Here T is a given finite time horizon, $\tau_{0:T} \triangleq (\tau_{0:T-1}, x_T)$, and $y_{0:T}$ denotes sequences of state-control pairs and measurement output from $t = 0$ to $t = T$, respectively. The joint probability density of the state-control trajectory $\tau_{0:T}$ and measurement trajectory $y_{0:T}$ is denoted as $p(\tau_{0:T}, y_{0:T}|\theta, \phi)$ and depends on the dynamic system parameter θ and control parameter ϕ. Further, $\mathbb{E}$, denotes the expected value of a random variable, when subscripted we highlight the measure, and, $L\left(\tau_{0:T}\right)$, is the standard cumulative cost defined as

$$L\left(\tau_{0:T}\right) \triangleq l_T(x_T) + \sum_{t=0}^{T-1} l_t(\tau_t) \tag{5}$$

The constant $\gamma \in \mathbb{R}_+$ in (4) is the risk-sensitivity parameter. Applying Taylor series expansion on the logarithm of the objective function in terms of L yields

$$-\frac{1}{\gamma}\log\left(\mathbb{E}\left[e^{-\gamma L}\right]\right) = \mathbb{E}\left[L\right] - \frac{\gamma}{2}\mathbb{V}\mathrm{ar}\left[L\right] + \mathcal{O}(\gamma \mathbb{V}\mathrm{ar}\left[L\right]) \tag{6}$$

where $\mathbb{V}\mathrm{ar}$ denotes the variance of a random variable and $\mathcal{O}$ denotes high order terms [3]. Therefore, the objective function (4) takes into account the expectation and variability of cost L, which is one attractive motivation for considering the objective function (4) to design an RSOC. In conclusion, remark that the RSOC problem (4) collapses to the risk-neutral objective when $\gamma \to 0$.

B. Identification subproblem

Note that to solve (4) it is required that θ is known. The second problem that we aim to treat is the estimation of parametric uncertainty. Therefore we adopt the probabilistic system identification framework where the identification problem is formulated as a Maximum Likelihood estimation problem [17]. The uncertain parameters of the dynamic system (1) are estimated by minimising the negative log-likelihood of the measurement output at all time steps with respect to θ [17].

$$\hat{\theta} = \arg\min_{\theta} -\log\left(p(y_{0:T}|\theta, \phi)\right) \tag{7}$$

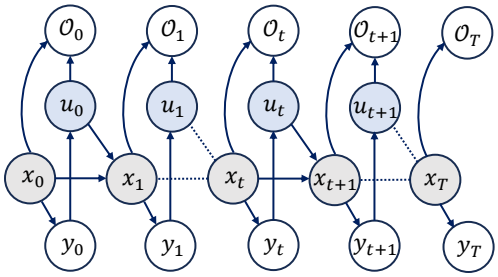

Fig. 2: PGM for RSOC

Some additional remarks are in place. As is well-known, treating problem (7) directly for any arbitrary nonlinear system is intractable [18]. Instead, in practice one relies on an iterative optimisation procedure known as the EM algorithm. Details are provided in section IV-B. Here it suffices to note that the EM algorithm produces a sequence of parameter estimates that converges to the optimal parameters by solving a series of easier surrogate problems. In a dynamic setting, the measurement sequence, $y_{0:T}$, increases with every sampling period. When the computational resources only allow a single iteration of the EM algorithm every period, a dynamic extension of the EM algorithm can be considered where the estimate from the previous iteration is used as prior for the next [19], [20]. This realises a dynamic identification procedure with manageable computational requirements.

C. Dual problem

We can now treat both problems separately. Ergo, we solve (7) in real-time for $\hat{\theta}_t$, then subsitute $\hat{\theta}_t$ for θ in (4) and solve for $\hat{\phi}_t$. This strategy would result in a control architecture without dual features for it implements the separation principle. Instead, we seek a tractable control architecture that exhibits dual features. To that end, we will embed the RSOC problem in the same formal framework as the identification problem and reformulate it as an inference problem. Specifically, this will allow us to embed (4) and (7) in an overarching problem statement that we can solve simultaneously. In the next section, we establish the probabilistic perspective on RSOC. In section IV we establish the overarching problem.

III. Probabilistic Perspective on RSOC

In this section, we discuss the representation of the RSOC problem as an inference problem using a probabilistic graph model (PGM). PGM refers to a type of probabilistic model where a graph is employed to represent the conditional dependency structure between random variables [21]. Fig. 2, shows the PGM for the RSOC. In this graphical model, a dummy random binary variable $\mathcal{O}_t \in \{0, 1\}$ at each time-step is introduced to indicate the notion of optimality or task fulfilment. Apart from their purpose and interpretation, these variables can be treated as regular observations. The state-control sequence is considered latent [1], [12].

When $\mathcal{O}_t = 1$ it is implied that time step t is optimal. To construct a convenient likelihood objective for the inference problem [22], we assume that the probability of being optimal at time t is proportional to an exponential utility transform of the immediate cost, $l_t(\tau_t)$ [9], [12], as follows

$$p(\mathcal{O}_t = 1|\tau_t) \propto \exp(-\gamma l_t(\tau_t)) \tag{8}$$

Based on Kolmogorov's probability axioms, the immediate cost, $l_t(\tau_t)$, needs to be defined by a positive structure for $\gamma > 0$. Hereafter, we unload the notation by simply writing $\mathcal{O}_t$ instead of $\mathcal{O}_t = 1$.

The following factorisation of the joint density can be obtained by considering factors in the PGM from Fig. 2.

$$\begin{aligned} p(\tau_{0:T}, \mathcal{O}_{0:T}, y_{0:T}|\theta, \phi) &= p(x_0)p(\mathcal{O}_T|x_T)p(y_T|x_T) \\ &\times \prod_{t=0}^{T-1} p(\mathcal{O}_t|\tau_t)p(y_t|x_t)p(x_{t+1}|\tau_t;\theta)p(u_t|y_t;\phi) \end{aligned} \tag{9}$$

where $p(x_{t+1}|\tau_t;\theta)$ indicates the state transition probability density parameterized by θ according to the dynamic model (1), $p(y_t|x_t)$ indicates the probability distribution of the measurement output according to the dynamic model (2), $p(u_t|y_t;\phi)$ represents the probability density of the controller (3) parameterized by ϕ, $p(\mathcal{O}_t|\tau_t)$ is defined in (8), $p(x_0)$ denotes the initial state probability, and $p(\mathcal{O}_T|x_T)$ and $p(y_T|x_T)$ refer to the terminal cost and the probability of the final state measurement, respectively.

Finally, we state the following theorem, inspired by [1], to cast the RSOC problem into a variational inference problem for controller design.

Theorem 1. *The minimization of the risk-sensitive objective function (4) w.r.t. controller parameter ϕ for given a parametric uncertainty vector θ is equivalent to the minimization of the negative log-likelihood (neg-log-likelihood) of the optimality variable at all time steps.*

$$\arg\min_{\phi} J_\gamma(\theta, \phi) = \arg\min_{\phi} -\frac{1}{\gamma}\log\left(p(\mathcal{O}_{0:T}|\theta, \phi)\right) \tag{10}$$

Proof. By substituting (8) in (9), and applying the property of exponents in multiplication expressions on its result, we have

$$p(\tau_{0:T}, \mathcal{O}_{0:T}, y_{0:T}|\theta, \phi) \propto p(\tau_{0:T}, y_{0:T}|\theta, \phi)\exp\left(-\gamma L\left(\tau_{0:T}\right)\right) \tag{11}$$

where $L\left(\tau_{0:T}\right)$ is introduced in (5), and $p(\tau_{0:T}, y_{0:T}|\theta, \phi)$ is specified by

$$p(x_0)p(y_T|x_T)\prod_{t=0}^{T-1} p(y_t|x_t)p(x_{t+1}|\tau_t;\theta)p(u_t|y_t;\phi) \tag{12}$$

By taking the integral of both sides of (11) w.r.t. all possible trajectories $\tau_{0:T}$ and $y_{0:T}$, we obtain

$$\begin{aligned} &p(\mathcal{O}_{0:T}|\theta, \phi) \\ &= \int p(\tau_{0:T}, \mathcal{O}_{0:T}, y_{0:T}|\theta, \phi)d\tau_{0:T}dy_{0:T} \\ &\propto \int p(\tau_{0:T}, y_{0:T}|\theta, \phi)\exp\left(-\gamma L\left(\tau_{0:T}\right)\right)d\tau_{0:T}dy_{0:T} \\ &= \mathbb{E}_{p(\tau_{0:T}, y_{0:T}|\theta,\phi)}\left[\exp\left(-\gamma L\left(\tau_{0:T}\right)\right)\right] \end{aligned} \tag{13}$$

Eventually, by taking the logarithm of (13) and multiplying $-1/\gamma$ by its result, the equivalency in (10) is established.

Note that the proportionality constant is the same for any trajectory $\tau_{0:T}$. □

In the next section, we show how both inference problems (10) and (7) for controller design and identifying the dynamic system can be combined as an AIF problem.

IV. Adaptive RSOC as Active Inference

In this section, we propose a tractable alternative for the adaptive dual control framework using an active inference framework to solve the adaptive RSOC problem.

Recall that AIF is a concept from theoretical neuroscience that is founded on the free energy principle (FEP) in the context of cognition and decision-making. Active inference aims to explain how biological or artificial agents, including the human brain, interact with their environment to actively reduce free energy, which includes minimizing the surprise or uncertainty of any event. Reducing surprisal or uncertainty for an agent occurs through making predictions based on internal models, updating them using sensory input, and taking actions to bring those predictions in line with sensory input [16].

To cast controller design (10) and system identification (7) as an active inference problem, we assess the following simultaneous surprisal minimization (the division by $\gamma > 0$ is neglected in the optimization process).

$$\min_{\phi,\theta} -\log\left(p(\mathcal{O}_{0:T}, y_{0:T}|\theta,\phi)\right) \tag{14}$$

A. Surprisal Minimization

We state the following lemma to explain the relationship between surprisal minimization (14) and the mentioned inference problems (10) and (7).

Lemma 1. *The neg-log-likelihood (14) can be decomposed in the following two ways. To simplify the notation, the subscript $0:T$ is eliminated for $\mathcal{O}$, y, and τ, and the conditioning on θ and ϕ is not characterized in the probability densities. $\mathbb{KL}$ denotes the standard Kullback–Leibler (KL) divergence for two probability distributions.*

$$\begin{aligned}
\text{(i)}\quad -\log\left(p(\mathcal{O},y)\right) &= -\log(p(\mathcal{O})) \underbrace{-\,\mathbb{E}_{p(\tau|\mathcal{O},y)}\left[\log(p(y|\tau))\right]}_{\text{Observational Exploration}} \\
&\quad \underbrace{+\,\mathbb{KL}\left[p(\tau|\mathcal{O},y)\parallel p(\tau|\mathcal{O})\right]}_{\text{Optimality Divergence}} \\
\text{(ii)}\quad -\log\left(p(\mathcal{O},y)\right) &= -\log(p(y)) \underbrace{-\,\mathbb{E}_{p(\tau|\mathcal{O},y)}\left[\log(p(\mathcal{O}|\tau))\right]}_{\text{Optimality Exploration}} \\
&\quad \underbrace{+\,\mathbb{KL}\left[p(\tau|\mathcal{O},y)\parallel p(\tau|y)\right]}_{\text{Observational Divergence}}
\end{aligned}$$

Proof. By expanding the right-hand side of (i) and using the Bayes' rule for $p(\mathcal{O}) = p(\tau,\mathcal{O})/p(\tau|\mathcal{O})$, we obtain

$$-\int_\tau p(\tau|\mathcal{O},y)\log\left(\frac{p(\tau,\mathcal{O})}{p(\tau|\mathcal{O})}p(y|\tau)\frac{p(\tau|\mathcal{O})}{p(\tau|\mathcal{O},y)}\right)d\tau \tag{15}$$

We know the fact that $\mathcal{O}$ and y are conditionally independent given τ, so $p(\tau,\mathcal{O},y) = p(\tau,\mathcal{O})p(y|\tau)$. The simplification of the expression in the parenthesis results

$$-\int_\tau p(\tau|\mathcal{O},y)\log\left(\frac{p(\tau,\mathcal{O},y)}{p(\tau|\mathcal{O},y)}\right)d\tau = -\log\left(p(\mathcal{O},y)\right) \tag{16}$$

The equality of (ii) can be proven in the same way. □

Considering the decomposition (i), we can interpret that minimising the neg-log-likelihood (14) w.r.t. ϕ to design the controller, in addition to optimizing the neg-log-likelihood (10), ensures that observations $y_{0:T}$ are as likely as possible under the state-action pairs $\tau_{0:T}$ while simultaneously minimizing the optimality divergence term, which means the goal is to keep the density $p(\tau|\mathcal{O},y)$ as close to $p(\tau|\mathcal{O})$ as possible while maximizing observational exploration. Effectively, the optimality divergence term acts as a regularizer to reduce the overfitting to any specific observation. Briefly, we try to optimize the neg-log-likelihood (10) while exploring the observation space y as little as possible to have a better convergence for the system identification. Similarly, we can explain the decomposition (ii) to interpret the system identification by minimizing the neg-log-likelihood (14) w.r.t. θ. Finding θ using (14), besides taking neg-log-likelihood (7) into account, considers the optimality exploration term to explore the optimality space while simultaneously minimizing the observational divergence term to avoid the overfitting.

Corollary 1. *We can conclude that concurrently minimizing the inference problem (14) w.r.t. θ and ϕ can be more valuable and effective to separately solve the problems (10) and (7) because, for finding each of θ and ϕ, we consider a regulated exploration in the space of the other one. Thus, balancing exploitation and exploration is inherent in our approach. So, we introduced a new framework as an adaptive dual control.*

B. Free Energy Minimization

Solving the optimization problem in (14) requires computing the model evidence $p(\mathcal{O},y) = \int_\tau p(\mathcal{O},y,\tau)d\tau$ which is analytically intractable [12], [16]. Therefore, we aim to derive a variational free energy bound ($\mathcal{F}$) for the surprisal minimization (14) according to [16]. To that end, consider the arbitrary *inference* distribution, $q(\tau_{0:T})$. Then we evaluate the Kullback-Leibler divergence between the *inference* distribution and the posterior distribution, $p(\tau_{0:T}|\mathcal{O}_{0:T},y_{0:T};\theta,\phi)$.

$$\begin{aligned}
0 &\le \mathbb{KL}[q(\tau_{0:T})\parallel p(\tau_{0:T}|\mathcal{O}_{0:T},y_{0:T};\theta,\phi)] \\
&= \mathbb{E}_{q(\tau_{0:T})}\left[\log(q(\tau_{0:T})) - \log(p(\tau_{0:T}|\mathcal{O}_{0:T},y_{0:T};\theta,\phi))\right] \\
&= \underbrace{\mathbb{E}_{q(\tau_{0:T})}\left[\log(q(\tau_{0:T})) - \log(p(\tau_{0:T},\mathcal{O}_{0:T},y_{0:T}|\theta,\phi))\right]}_{\mathcal{F}_{0:T}(q,\theta,\phi)} \\
&\quad + \log(p(\mathcal{O}_{0:T},y_{0:T}|\theta,\phi))
\end{aligned} \tag{17}$$

Here the variational free energy is defined as the Kullback-Leibler divergence between our inference distribution and the joint density. Due to the positive definiteness of the Kullback-Leibler divergence, the variational free energy bounds the surprisal as follows

$$-\log(p(\mathcal{O}_{0:T},y_{0:T}|\theta,\phi)) \le \mathcal{F}_{0:T}(q,\theta,\phi) \tag{18}$$

Thus instead of minimizing the surprisal, (14), we can minimize the free energy which, by construction, proves to be an upper bound on the former. Further remark that the free energy depends on the arbitrary inference distribution. The gap between surprisal and free energy is minimized if the inference distribution coincides with the posterior.

This invites an iterative procedure better known as the EM algorithm. The EM algorithm is an iterative method that operates by alternating between two main steps: first, finding the lowest upper bound for the neg-log-likelihood function through inferring the density of latent variables (known as the E-step); and then, minimizing this bound w.r.t. the unknown parameters (known as the M-step). The repetition of these E-step and M-step iterations leads to convergence such that it is guaranteed to converge to a local optimum [21].

- **E-step:** The inequality of (18) turns to equality when the KL divergence term in (17) takes zero or equivalently $q(\tau_{0:T}) = p(\tau_{0:T}|\mathcal{O}_{0:T}, y_{0:T}; \theta, \phi)$. Consequently, we need to compute the smoothed density of the state-control pairs given the estimated parameters from M-step [12]. To this end, any smoothing algorithms in [23] for nonlinear dynamic systems can be applied, such as the Extended or Unscented Rauch-Tung-Striebel (RTS) smoothing algorithms.
- **M-step:** Using a given smoothed density of the state-control pairs from E-step and substituting the joint probability of (9) in the variational free energy bound (17), we retrieve the subsequent optimization problem to find the parameters θ and ϕ.

$$\arg\min_{\theta,\phi} -\mathbb{E}_{q(\tau_{0:T})}\Bigg[\log\Bigg(p(x_0)p(\mathcal{O}_T|x_T)p(y_T|x_T)\times \prod_{t=0}^{T-1} p(\mathcal{O}_t|\tau_t)p(y_t|x_t)p(x_{t+1}|\tau_t;\theta)p(u_t|y_t;\phi)\Bigg)\Bigg] \tag{19}$$

C. Receding Horizon Strategy

The problem proposed in (14) cannot be applied in a real-time setting. To that end, we propose a generalization with a receding prediction horizon [24]. To incorporate the receding horizon strategy into the EM procedure over the fixed time horizon H, instead of (14), we consider the adapted surprise $p(\mathcal{O}_{0:t+H}, y_{0:t})$. Using Bayes' rule it becomes $p(\mathcal{O}_{t+1:t+H}|\mathcal{O}_{0:t}, y_{0:t})p(\mathcal{O}_{0:t}, y_{0:t})$. It decomposes the variational upper bound optimization mentioned in (17) and (18) into two time-windows: from 0 to t to update the system parameter estimation relating to $p(\mathcal{O}_{0:t}, y_{0:t})$ and from $t+1$ to $t+H$ to compute the controller parameters relating to $p(\mathcal{O}_{t+1:t+H}|\mathcal{O}_{0:t}, y_{0:t})$. By considering the new surprisal in the free energy bound (17), we also need to infer the prediction of the future observations of the system for free energy minimization (19), i.e., we should infer the smoothed density $\hat{q}(\tau_{0:t+H}, y_{t+1:t+H}|y_{0:t}, \mathcal{O}_{0:t+H}; \hat{\theta}, \hat{\phi})$ at each time step which can be obtained using Bayes' rule as

$$\hat{q}(\tau_{0:t+H}|y_{0:t}, \mathcal{O}_{0:t+H}; \hat{\theta}, \hat{\phi})p(y_{t+1:t+H}|\tau_{t+1:t+H}; \hat{\theta}, \hat{\phi}) \tag{20}$$

By taking the logarithm of the product operator and marginalizing the trajectories in the expectation of (19) and using (20), we decompose and simplify (19) into two time-windows as

$$\arg\min_{\theta,\phi} J_{\text{past}}(\theta) + J_{\text{future}}(\phi) \tag{21}$$

with

$$J_{\text{past}}(\theta) = -\sum_{k=0}^{t} \mathbb{E}_{\hat{q}(\tau_{k+1},\tau_k)}\Big[\log\left(p(x_{k+1}|\tau_k;\theta)\right)\Big] \tag{22}$$

$$J_{\text{future}}(\phi) = -\sum_{k=t+1}^{t+H-1} \mathbb{E}_{\hat{q}(\tau_k)p(y_k|x_k)}\Big[\log\left(p(u_k|y_k;\phi)\right)\Big], \tag{23}$$

and where $\hat{q}(\cdot|y_{0:t}, \mathcal{O}_{0:t+H}; \hat{\theta}, \hat{\phi})$ is the smoothed density mentioned in (20). In the past time of t, we applied the previously optimized controller parameters and observed the past measurements of the system. Therefore, the past terms of (19) do not participate in the optimization procedure w.r.t. ϕ. Likewise, we employ the estimated system parameter θ to predict the future of the controlled system. Hence, the future terms of (19) are not engaged in the optimization procedure w.r.t. θ. We acknowledge that there are alternative optimization strategies to be explored however in this work we have restricted to that described here.

Solving the optimization problem (21) in each time step introduces a novel adaptive model predictive control (AMPC) algorithm for stochastic nonlinear systems with the risk-sensitive optimality criterion. Also, employing the shifted or receded fixed prediction horizon over time can enable our framework to be applied in an infinite-horizon framework.

D. Special case

As a special case we consider a dynamic system (1) and controller (3) that are linearly parameterized by basis functions $\psi(\tau_t) \in \mathbb{R}^{n_x \times n_\theta}$ and $\sigma(y_t) \in \mathbb{R}^{n_u \times n_\phi}$, respectively. Also, we assume a well-known quadratic cost as, $l_t(\tau_t)$, which will be encoded into the probability of the optimality variable $\mathcal{O}_t$. Therefore, we have

$$\begin{aligned} f(\tau_t;\theta) = \psi(\tau_t)\theta \Rightarrow & p(x_{t+1}|\tau_t;\theta) = \mathcal{N}(x_{t+1}|\psi(\tau_t)\theta, V) \\ K(y_t;\phi) = \sigma(y_t)\phi \Rightarrow & p(u_t|y_t;\phi) = \mathcal{N}(u_t|\sigma(y_t)\phi, \Sigma) \\ & p(\mathcal{O}_t = 1|\tau_t) \propto \mathcal{N}(z_t = z_t^*|\tau_t, \Gamma^{-1}) \end{aligned} \tag{24}$$

where $z_t \in \mathbb{R}^{(n_x+n_u)}$ denotes the measurement variable relating to the optimality variable $\mathcal{O}_t$. We consider $\Gamma \triangleq \gamma \mathrm{Q}$ in which $\mathrm{Q} \in \mathbb{R}^{(n_x+n_u)\times(n_x+n_u)}$ is a positive-definite weighting matrix for a standard quadratic cost. Applying the optimization problem (21) on the linearly parameterized expressions mentioned in (24) yields the following two unconstrained quadratic programming w.r.t. θ and ϕ

$$\{\hat{\theta}^i, \hat{V}^i\} = \arg\min_{\theta,V} \sum_{k=0}^{t} \mathbb{E}_{\hat{q}^i(\tau_{k+1},\tau_k)}\left[(x_{t+1} - \psi(\tau_t)\theta)^\top V^{-1}(x_{t+1} - \psi(\tau_t)\theta)\right] \tag{25}$$

and

$$\{\hat{\phi}^i, \hat{\Sigma}^i\} = \arg\min_{\phi,\Sigma} \sum_{k=t+1}^{t+H-1} \mathbb{E}_{\hat{q}^i(\tau_k)p(y_k|x_k)}\left[(u_t - \sigma(y_t)\phi)^\top \Sigma^{-1}(u_t - \sigma(y_t)\phi)\right] \tag{26}$$

where i indicates the iteration index for the EM algorithm iterations and $\hat{q}^i(\cdot|y_{0:t}, \mathcal{O}_{0:t+H}; \hat{\theta}^{i-1}, \hat{\phi}^{i-1})$ refers to the smoothed density computed using estimated parameters from the previous iteration. For each time step, the optimization problems (25) and (26) are computed iteratively until convergence. Finally, the whole introduced algorithm is summarized in Algorithm 1 and shown in Fig. 1.

Algorithm 1: Adaptive RSOC

Input: Γ, $\psi(.)$, $\sigma(.)$, H, m_{x_0}, P_{x_0}, and N.
Output: ϕ_t^* and Σ_t^*.
for t **do** `// Real-time loop`
  $\hat{\theta}^0, \hat{V}^0, \hat{\phi}^0, \hat{\Sigma}^0 \leftarrow \theta_{t-1}^*, V_{t-1}^*, \phi_{t-1}^*, \Sigma_{t-1}^*$
  for $i = 1 : N$ **do** `// EM iteration`
    $x_0^p = y_t$
    for $k = 0 : H-1$ **do** `// Prediction`
      $y_k^p \sim \mathcal{N}(y_k^p | x_k^p, W)$
      $u_k^p \sim \mathcal{N}(u_k^p | \sigma(y_k^p)\hat{\phi}^{i-1}, \hat{\Sigma}^{i-1})$
      $x_{k+1}^p \sim \mathcal{N}(x_{k+1}^p | \psi(\tau_k^p)\hat{\theta}^{i-1}, \hat{V}^{i-1})$
      Store (u_k^p, y_k^p) in $\mathcal{D}_P$
    E-Step: Apply Extended/unscented RTS smoothing algorithm on $\mathcal{D}_R \cup \mathcal{D}_P$ to find $\hat{q}^i$
    M-Step: Update $\hat{\theta}^i$, $\hat{V}^i$, $\hat{\phi}^i$, and $\hat{\Sigma}^i$ using (25) and (26)
  $\theta_t^*, V_t^*, \phi_t^*, \Sigma_t^* \leftarrow \hat{\theta}^N, \hat{V}^N, \hat{\phi}^N, \hat{\Sigma}^N$
  $u_t \sim \mathcal{N}(u_t | \sigma(y_t)\phi_t^*, \Sigma_t^*)$
  Execute u_t and observe y_{t+1} from the dynamic system
  Store (u_t, y_t) in $\mathcal{D}_R$

V. Simulation Example

Consider the following nonlinear dynamic system derived from Duffing's equation which can model the behaviour of a noisy and unstable mass-spring-damper system with a nonlinear hardening spring.

$$\ddot{p} + b\dot{p} + k(1 + a^2p^2)p + v_t = u \tag{27}$$

where p, $\dot{p}$, and $\ddot{p}$ indicate displacement, velocity, and acceleration, respectively, $b < 0$ and $k > 0$ are related to the negative damping coefficient and the spring constant normalized by mass, and a is the non-linearity constant for the hardening spring. v_t is a zero-mean Gaussian noise.

We considered 60% parametric uncertainty on the nominal values for the simulations. Table I shows the nominal and perturbed parameter values.

The following proportional-integrator-derivative (PID) controller is applied to the dynamic system (27).

$$\begin{aligned} \dot{x}_c &= p + \dot{p} \\ u &= -K_1 p - K_2 \dot{p} - K_3 x_c \end{aligned} \tag{28}$$

in which x_c is the integrator state of the controller which will be augmented in the state vector of the dynamic system, K_1, K_2, and K_3 are the controller gains supposed to be designed. After time-discretization of (27) and (28) with sampling time $t_s = 0.1$ [s] and defining the time-discretized state vector as $x_t \triangleq [p_t, \dot{p}_t, x_{c_t}]^\top$, we can achieve the linear

TABLE I: Nominal and Perturbed Parameter Values

Parameter	b [1/s]	k [1/s^2]	a
Nominal Value	-0.1	1	0.1
Perturbed Value	-0.16	0.4	0.04

TABLE II: Performance Value for Different Settings

Variation of H (N is fixed and $N = 30$)	$H = 30$	$H = 50$	$H = 80$
Cost Function J	5.26	3.72	3.44
Variation of N (H is fixed and $H = 50$)	$N = 20$	$N = 30$	$N = 60$
Cost Function J	4.05	3.72	3.41

parameterization form in (24). We solve the Adaptive RSOC problem (14) using Algorithm 1 to design the controller gains K_1, K_2, and K_3 in the presence of parametric uncertainty with the following settings

$$\begin{aligned} \mathrm{Q} &= \mathrm{diag}([0.25, 0.1, 0.1, 0.1]) \\ \gamma &= 0.25 \\ m_{x_0} &= [5, 2, 0]^\top \\ V &= W = P_{x_0} = 10^{-4}\mathbb{I}_3 \\ \Sigma &= 10^{-4} \end{aligned} \tag{29}$$

where $\mathbb{I}_3$ is a 3×3 identity matrix and diag denotes the diagonal matrix operator.

The performance of the proposed algorithm 1 for different horizon lengths and different numbers of iterations of the EM algorithm are reported in Table II which demonstrates that by increasing the horizon length or the number of iterations for the EM algorithm, we can achieve more optimal performance.

We applied the LEQG controller [3] on the linearized dynamic system (27) to design the PID controller gains. If the LEQG controller is aware of the uncertainty values on the system's parameters, the cost function value and the states and control input trajectories are revealed in Table III and Fig. 3 as "Optimal LEQG". In this case, we have an optimal controller for the system perturbed by the parametric uncertainty and we can compare different methods and settings with it to find how much they are far from the optimal solution. Furthermore, we designed the LEQG controller using the nominal values of the system's parameters and applied it to the perturbed system. Its results are depicted as "Perturbed LEQG" in Table III and Fig. 3 such that it cannot preserve the stability of the system in the presence of uncertainty.

Also, we separately solved the system identification inference problem (7) and the controller design inference problem (10) at each time step in a receding horizon strategy. In this approach, the EM algorithm and finding the extended/unscented RTS smoothing density are separately carried out for both inference problems. Table III and Fig. 3 signify our algorithm has more optimal results than separately conducting system identification and controller design ("SysID+ContDesign") because as we showed in Subsection IV-A, considering minimization of the inference problem (14) can improve the balance between exploration and exploitation efficiently. Ergo, the proposed control architecture exhibits dual features.

TABLE III: Performance of Different Methods

Method	Cost Function
Optimal LEQG	2.84
Our Algorithm ($H = 50$, $N = 60$)	3.41
SysID+ContDesign ($H = 50$, $N = 60$)	6.99
Perturbed LEQG	10.85

VI. Conclusion

The main goal of the current study was to merge system identification and controller design within an active inference framework to solve the RSOC problem whilst exhibiting dual features. This paper employed the EM algorithm, combining the RTS smoother in the E-Step and predictive parameter updates in the M-Step, to effectively design an adaptive controller for uncertain stochastic systems. The proposed approach yields a natural balance between exploitation and exploration and results in a computationally efficient quadratic programming solution. Numerical simulations confirmed the performance of the proposed algorithm framework. Future efforts will focus on considering other risk metrics like CVaR and encoding safety constraints within the proposed algorithm. It would also be interesting to consider a broader class of stochastic systems like non-Gaussian systems.

Acknowledgment

This work was supported by the Research Foundation Flanders (FWO) under SBO grant n S007723N and the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

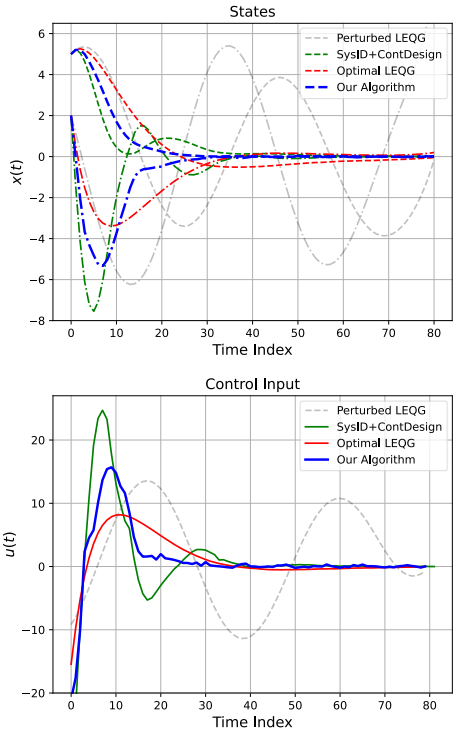

Fig. 3: Upper figure: state vector signals p_t [m] (dashed curves) and $\dot{p}_t$ [m/s] (dash-dotted curves) and lower figure: control input signal u_t [N/kg] (solid curves) for different approaches.

References

[1] E. Noorani and J. S. Baras, "A probabilistic perspective on risk-sensitive reinforcement learning," in *Proceedings of the 2022 American Control Conference (ACC)*, 2022, pp. 2697–2702.

[2] B. Hammoud, A. Jordana, and L. Righetti, "iRiSC: Iterative risk sensitive control for nonlinear systems with imperfect observations," in *Proceedings of 2022 American Control Conference (ACC)*, 2022, pp. 3550–3557.

[3] P. Whittle, *Optimal Control: Basics and Beyond*. Chichester, United Kingdom: John Wiley & Sons, 1996.

[4] Z. Wang, O. So, K. Lee, and E. A. Theodorou, "Adaptive risk sensitive model predictive control with stochastic search," in *Proceedings of 3rd Learning for Dynamics and Control Conference*. PMLR, 2021, pp. 510–522.

[5] Y. Fei, Z. Yang, Y. Chen, Z. Wang, and Q. Xie, "Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 384–22 395, 2020.

[6] Y. Fei, Z. Yang, and Z. Wang, "Risk-sensitive reinforcement learning with function approximation: A debiasing approach," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 3198–3207.

[7] A. Feldbaum, "Dual control theory, Part I," *Automation and Remote Control*, vol. 21, no. 9, pp. 874–880, 1961.

[8] A. FeldbauM, "Dual control theory, Part II," *Automation and Remote Control*, vol. 21, no. 11, p. 1033–1039, 1961.

[9] S. Levine, "Reinforcement learning and control as probabilistic inference: Tutorial and review," *arXiv preprint, arXiv:1805.00909*, 2018.

[10] J. Watson, H. Abdulsamad, and J. Peters, "Stochastic optimal control as approximate input inference," in *Proceedings of 3rd Conference on Robot Learning (CoRL)*. PMLR, 2019, pp. 697–716.

[11] J. Watson and J. Peters, "Advancing trajectory optimization with approximate inference: Exploration, covariance control and adaptive risk," in *Proceedings of 2021 American Control Conference (ACC)*, 2021, pp. 1231–1236.

[12] S. P. Q. Syed and H. Bai, "Parameterized input inference for approximate stochastic optimal control," in *Proceedings of the 2023 American Control Conference (ACC)*, 2023, pp. 2574–2579.

[13] M. Toussaint, "Robot trajectory optimization using approximate inference," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009, pp. 1049–1056.

[14] K. Rawlik, M. Toussaint, and S. Vijayakumar, "An approximate inference approach to temporal optimization in optimal control," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[15] K. Friston, S. Samothrakis, and R. Montague, "Active inference and agency: optimal control without cost functions," *Biological Cybernetics*, vol. 106, pp. 523–541, 2012.

[16] L. Da Costa, T. Parr, N. Sajid, S. Veselic, V. Neacsu, and K. Friston, "Active inference on discrete state-spaces: a synthesis," *Journal of Mathematical Psychology*, vol. 99, p. 102447, 2020.

[17] T. B. Schön, A. Wills, and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.

[18] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4, no. 4.

[19] J. Daunizeau, K. J. Friston, and S. J. Kiebel, "Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models," *Physica D: nonlinear phenomena*, vol. 238, no. 21, pp. 2089–2118, 2009.

[20] W. Kouw, A. Podusenko, M. Koudahl, and M. Schoukens, "Variational message passing for online polynomial NARMAX identification," in *Proceedings of the 2022 American Control Conference (ACC)*, 2022, pp. 2755–2760.

[21] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. Cambridge, MA, USA: MIT press, 2023.

[22] J. Peters and S. Schaal, "Reinforcement learning by reward-weighted regression for operational space control," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007, p. 745–750.

[23] S. Särkkä and L. Svensson, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2023, vol. 17.

[24] J. Mattingley, Y. Wang, and S. Boyd, "Receding horizon control," *IEEE Control Systems Magazine*, vol. 31, no. 3, pp. 52–65, 2011.