Application of the Free Energy Principle to Estimation and Control

Thijs van de Laar*, Ayça Özçelikkale†, and Henk Wymeersch‡

*Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands
†Dept. of Electrical Engineering, Uppsala University, Sweden
‡Dept. of Electrical Engineering, Chalmers University of Technology, Sweden

December 31, 2021

Abstract

Based on a generative model (GM) and beliefs over hidden states, the free energy principle (FEP) enables an agent to sense and act by minimizing a free energy bound on Bayesian surprise. Inclusion of prior beliefs in the GM about desired states leads to active inference (ActInf). In this work, we aim to reveal connections between ActInf and stochastic optimal control. We reveal that, in contrast to standard cost and constraint-based solutions, ActInf gives rise to a minimization problem that includes both an information-theoretic surprise term and a model-predictive control cost term. We further show under which conditions both methodologies yield the same solution for estimation and control. For a case with linear Gaussian dynamics and a quadratic cost, we illustrate the performance of ActInf under varying system parameters and compare to classical solutions for estimation and control.

Index terms — *Active Inference, Stochastic Optimal Control, Message Passing, Factor Graphs*

*TvdL acknowledges the support from GN Hearing A/S and the Netherlands Organization for Scientific Research, project number 13925.

†AÖ acknowledges the support from the Swedish Research Council, Grant 2015-04011.

‡HW acknowledges the support from the Swedish Research Council, Grant 2018-03701. The authors thank Themistoklis Charalambous for valuable comments and Magnus Koudahl for the stimulating discussions.

1 Introduction

Bayesian graphical models (BGMs) constitute an important family of tools in signal processing, as they allow learning of models as well as inference of hidden states in a unified way, often with low complexity. BGMs have been widely applied for a wide variety of estimation and detection problems in signal processing [1, 2], with applications that include sensor networks [3], surveillance [4], and information-seeking control [5]. They also naturally unify several standard methods from statistical estimation theory, such as the forward-backward algorithm [6], the Kalman filter [7], the particle filter [8], and the Viterbi algorithm [7].

Beyond learning, estimation and detection, BGMs have also found applications in stochastic control problems, which involve not only estimation of the state of a system, but also determination of suitable control actions in the presence of uncertainty. Applications include control of vehicles, robots, factories, or teams of agents. Often, the control problem and inference/estimation problem are considered separate, whereby the controller assumes an estimate of the state and the inference occurs without knowledge of future control. Such a separation principle is only valid in certain cases, such as the celebrated linear quadratic Gaussian (LQG) control [9]. Over the past decades, several approaches have been proposed to unify inference and control [10–16], largely based on BGMs. The core ideas of these approaches can already be found in the early work [10], which posed the role of a controller as follows: *"A controller of a stochastic system 'shapes' the joint pdf describing the closed-loop behaviour. The 'optimal' controller should make this joint pdf as close as possible to a desired pdf."* With this in mind, [10] poses an ideal state distribution and control distribution, after which an optimized controller can be found be minimizing a Kullback-Leibler divergence (KLD). In terms of mathematical tractability, an important improvement was the use of Bayesian graphical models [2], which led to the methods in [12, 13, 16]. In [12], a similar idea as [10] was proposed, which allowed the formulation of control cost as a KLD, which could be solved by approximate inference on the corresponding graphical model. In [13], a linear transformation of the control cost function is replaced by a log-likelihood function and an optimized controller is found by an expectation-maximization procedure over the corresponding factor graph. A similar idea was introduced in [14, 15] where an artificial observation and associated likelihood was introduced so that state trajectories with highest posterior probability also have lowest associated control cost. In [13, 15] controllers similar to LQG were found. In [16] the LQG control problem was targeted specifically, and under perfect knowledge of the current state, the exact LQG controller was recovered by an inference-based controller. It should be noted that some of the above works aim to find a policy (i.e. a mapping from state estimate to control) while others aim to determine an optimal control sequence.

More generally, stochastic optimal control problems have been solved using a diverse range of approaches, where *model-predictive control (MPC)* [17] and *reinforcement learning (RL)* [18] are arguably the most prominent approaches.

When a model of the system is available, the control problem becomes a Markov decision process, which can, in principle, be solved through dynamic programming [19]. If no model is available, RL can provide model-free solutions that learn state-action mappings from interactions with the system [20]. Recently, there has been work combining these two approaches, originating either from the control theory community [21] or the computer science community [22].

In addition to MPC and RL, a third and more recent path is the *free energy principle (FEP)*, which originates from cognitive neuroscience as a way to explain biological behavior [23, 24]. The main hypothesis is that agents (i) internalize a generative model (GM) of the system, and (ii) perceive and act in such a way as to minimize a free-energy bound on surprise relative to the GM. Interestingly, free-energy minimization is a concept that is also used in RL to encourage exploration and model building [25]. Objective functions for any kind of system or application can be included in the GM in the form of a goal prior, which results in formulations of active inference (ActInf) [26]. Despite a large number of publications in the ActInf field including applications in robotics [27] and synthesis with reinforcement learning [28], there have been only few efforts, e.g., [16, 29–33], to apply it to more traditional stochastic control settings, such as linear quadratic Gaussian (LQG) control.

In this paper, our main aim is to reveal the connections between inference and control over BGM from an ActInf perspective. Specifically, we have the following contributions:

1. We propose an ActInf joint inference and control formulation that casts the control problem as an inference problem and explicitly encodes the control cost in the FEP framework;

2. We show under which conditions the ActInf-based joint inference and control method yields the solution to the original stochastic optimal control problem;

3. We prove that LQG can be expressed as a special case of the proposed ActInf joint inference and control method.

The article is structured as follows: In the remainder of this section, we provide an overview of the notation. Sec. 2 introduces the model and optimal control objective. Sec. 3 introduces the ActInf objective and further notation related to probabilistic model formulations. Sec. 4 formally relates the objective function of ActInf with stochastic optimal control. Sec. 5 then applies these formulations to a LQG control problem, which is illustrated by the numerical results in Sec. 6. We conclude with Sec. 7.

Notation

We write a sequence of variables as $s_{t_1:t_2} = \{s_{t_1}, \dots, s_{t_2}\}$. At any current time t, we consider a sequence of states, observations and controls as $x = x_{0:t+T}$, $y = y_{1:t+T}$, $u = u_{0:t+T-1}$, with $x_k \in \mathbb{R}^{n_x}$, controls $u_k \in \mathbb{R}^{n_u}$, and observations

Table 1: Common notations for distributions and functions.

	Short Description	*Normalized*
p	joint distribution	yes
p_t	generative model at time t	yes
$\tilde{p}$	goal priors	yes
f_t	goal-constrained generative model	no
p_p	posterior distribution of hidden variables	yes
q	belief / variational posterior	yes
π_t	stochastic policy mapping	yes

$y_k \in \mathbb{R}^{n_y}$ respectively, with a time horizon of T time-steps into the future. Note that the start and the end points between the state, observation and control sequence differ slightly. When explicitly required, we denote the realizations of the random variables, such as (past) observed values, estimates and performed actions, by a bold script.

In order to easily distinguish between the past and future variables, we adopt the following convention: we divide the observations y into past (including present) variables $\underline{y}_t = y_{1:t}$ and future variables $\overline{y}_t = y_{t+1:t+T}$. Similarly, the state sequence x consists of $\underline{x}_t = x_{0:t}$ and $\overline{x}_t = x_{t+1:t+T}$. The control sequence u consists of $\underline{u}_t = u_{0:t-1}$ and $\overline{u}_t = u_{t:t+T-1}$ (with present control included). For notational convenience, we drop the dependence on the current time t. For instance, we use $\overline{x}$ instead of $\overline{x}_t$. We use $s_{\setminus t}$ to denote the sequence obtained by removing s_t from s.

Similar to the notation for the sequences, some of the probability density functions (pdfs) are expressed using the notation $\underline{p}(.)$ and $\overline{p}_t(.)$ to emphasize functions of past and future variables, respectively. As usual, marginal and conditional pdfs associated with a given joint pdf are denoted using the same letter/subscript. For instance, the marginal obtained by marginalizing (i.e., integrating) $p_a(s_1, s_2)$ over s_2 is denoted by $p_a(s_1) = \int p_a(s_1, s_2)\, \mathrm{d}s_2$. To avoid clutter, we drop the distribution arguments (i.e., we write p_a instead of $p_a(s_1, s_2)$) whenever these dependencies are clear from the context.

2 System Model

2.1 Dynamical System Model

We consider the following dynamical system with the state-space model (SSM):

$$x_{t+1} \sim p(x_{t+1}|x_t, u_t), \quad t \geq 0\,, \tag{1a}$$

$$y_t \sim p(y_t|x_t), \quad t \geq 1\,, \tag{1b}$$

where $x_0 \sim p(x_0)$ and $u_0 = 0$. Using the system definition in (1), the probabilistic system model for the state and outcome sequence for a given control

sequence over a time window of $k \in [0, t+T]$ can be expressed as follows

$$p(y, x|u) = p(x_0) \prod_{k=0}^{t+T-1} p(y_{k+1}|x_{k+1}) p(x_{k+1}|x_k, u_k). \tag{2}$$

At time t, we have the probabilistic system model

$$p_t(y, x|u) = \frac{p(\underline{y}_t = \underline{\boldsymbol{y}}_t, \overline{y}_t, x|\underline{u}_t = \underline{\boldsymbol{u}}_t, \overline{u}_t)}{\int p(\underline{y}_t = \underline{\boldsymbol{y}}_t, \overline{y}_t, x|\underline{u}_t = \underline{\boldsymbol{u}}_t, \overline{u}_t) \, \mathrm{d}\overline{y}_t \, \mathrm{d}x}, \tag{3}$$

where $\underline{y}_t$ and $\underline{u}_t$ are set to their realizations ($\underline{\boldsymbol{y}}_t$ and $\underline{\boldsymbol{u}}_t$). We will generally omit the explicit dependence on $\underline{\boldsymbol{y}}_t$ and $\underline{\boldsymbol{u}}_t$ and instead rely on the sub-script t to indicate that past controls and observations are fixed in p_t. Since p_t is obtained by plugging in the values of the realizations, we re-normalize. An overview of the common distributions used in this paper together with their normalization status is provided in Table 1.

2.2 Control Objective

We consider stochastic policy mappings in the form of $\pi_k(u_k|y_{1:t})$ from the set of measurements (up to the current time t) to the control at time k where $k \in [t, t+T]$. The objective is to find the mappings π_k that minimize the expected cost $\mathcal{J}_t$ over current and future states x_k and controls u_k, defined as:

$$\mathcal{J}_t = \sum_{k=t}^{t+T} \mathbb{E}_{p_t, \pi_k} \left[\ell_k(x_k, u_k) \right] , \tag{4}$$

where p_t is the probabilistic system model as expressed in (3), and $\ell_k(x_k, u_k) \geq 0$ is the cost function at time-step k that encodes the cost of being in state x_k and applying the control u_k. The realization for the current control (action) is then determined using the stochastic policy mapping π_t.

In particular, the control is $u_t = \boldsymbol{u}^*_{t,\pi}$, where

$$\boldsymbol{u}^*_{t,\pi} = g(\pi^*_t) \tag{5}$$

with

$$\pi^*_t = \arg\min_{\pi_t} \mathcal{J}_t \tag{6}$$

and where $g(\cdot)$ represents the mapping from the probability distribution to a single action $\boldsymbol{u}_t$, which can be chosen, for instance, as the mean or the mode of π^*_t or as a sample (i.e. realization) from π^*_t [18,30,34,35]. This article considers a sliding horizon, i.e., after taking the action at the current time instant t and obtaining the next observation, the stochastic policies are again determined by looking T steps ahead.

Example: A typical cost function is the quadratic cost:

$$\ell_k(x_k, u_k) = \ell(x_k, u_k) = \tfrac{1}{2} x_k^{\mathrm{T}} Q x_k + \tfrac{1}{2} u_k^{\mathrm{T}} R u_k \, , \tag{7}$$

for $Q \in \mathbb{R}^{n_x \times n_x}, Q \succeq 0$ and $R \in \mathbb{R}^{n_u \times n_u}, R \succ 0$.

3 Active Inference

In this section, we describe the ActInf approach and the FEP. The concepts and approaches in this section have similarities to the control as inference literature [12, 13, 16], but are here presented from the ActInf perspective. The main idea of ActInf is that at each time t, the controller minimizes the free energy functional $\mathcal{F}_t[q]$, defined as [23] $\mathcal{F}_t[q] = \mathrm{D}[q||f_t]$, where $\mathrm{D}[q||f_t]$ is the Kullback-Leibler divergence, q represents a variational distribution (the optimization variable) and f_t represents the known generative model p_t with substituted observations or with modifications with more general constraints. Each of these concepts will now be explained in detail. Minimization of the free energy, and the well-established framework of minimization of Bayesian surprise are closely connected. We further discuss this relationship in Remark 1.

3.1 Generative Model and Goal Priors

3.1.1 The Generative Model

The notation introduced earlier allows us to concisely write the system model at time t (3) in terms of a past and a future contribution:

$$p_t(y, x|u) = p_t(\overline{y}, \underline{y}, \overline{x}, \underline{x}|\overline{u}, \underline{u}), \tag{8a}$$

$$= p_t(\underline{y}, \underline{x}|\overline{u}, \underline{u})\, p_t(\overline{y}, \overline{x}|\underline{y}, \underline{x}, \overline{u}, \underline{u}), \tag{8b}$$

$$= \underline{p}_t(\underline{y}, \underline{x}|\underline{u})\, \overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})\,. \tag{8c}$$

We note that the first factor depends on past controls and the second on the future controls. Both factors condition on the controls, and p_t does not incorporate the control cost $\mathcal{J}_t$.

3.1.2 Goal Priors

The designer of the agent should govern the system behaviour towards desirable system states, e.g. the exit of a maze. In order to achieve this, ActInf introduces the concept of a *prior belief on the future outcomes* [36–38] (or equivalently referred as a goal prior) which constrains the system model (2). This goal prior is set by the designer of the agent and encodes the future states that are desirable, in other words the states that we want to be *unsurprising* for the agent. Actions selected as a result of free energy minimization will then move the agent as close as possible to these unsurprising (desired) states. A goal prior is added as an additional factor to the system model description [37,38], leading to the goal-constrained (unnormalized) GM:

$$f_t(y, x, \overline{u}|\underline{u}) = \underbrace{p_t(y, x|u)}_{\substack{\text{generative}\\ \text{model}}} \underbrace{\tilde{p}(\overline{y}, x_t, \overline{x}, \overline{u})}_{\text{goal prior}}\,. \tag{9}$$

In order to relate the goal prior to the traditional control cost, a natural choice is:

$$\tilde{p}(\overline{y}, x_t, \overline{x}, \overline{u}) = \frac{1}{\Gamma} \exp\left(-\lambda \sum_{k=t}^{t+T} \ell_k(x_k, u_k)\right), \tag{10}$$

where Γ the the normalization constant and $\lambda \geq 0$ is the scaling factor. For the quadratic cost of (7) the goal prior factors into independent Gaussians (i.e., consists of factors in the form $\propto \exp(-\lambda \frac{1}{2} x_k^{\mathrm{T}} Q x_k) \exp(-\lambda \frac{1}{2} u_k^{\mathrm{T}} R u_k)$), where the weighting matrices Q and R take the role of (scaled) precisions. A related probabilistic approach is described in [15], where a binary reward is defined by using a cost function.

3.2 Free Energy Objective

Consider the latent (hidden) variables at time t: $\overline{y}, x, \overline{u}$. Note that the state sequence x is unknown for both the future and the past, whereas for the observations and the controls, only the future variables are unknown. Let us consider a variational posterior distribution $q(\overline{y}, x, \overline{u}|\underline{y}, \underline{u})$ defined over the latent variables. Here, the label *variational* refers to the fact that the objective function (11) is optimized by *variations* in the conditional [39]. Note that $q(\overline{y}, x, \overline{u}|\underline{y}, \underline{u})$ is a posterior conditioned on the past observations and controls. To avoid notational clutter, we adopt a mainstream notational convention in probabilistic inference where conditioning is dropped from the variational posterior distribution, and represent $q(\overline{y}, x, \overline{u}|\underline{y}, \underline{u})$ as $q(\overline{y}, x, \overline{u})$ or simply as q.

The free energy $\mathcal{F}_t[q]$ is defined as follows [23]:

$$\mathcal{F}_t[q] = \mathrm{D}[q||f_t]\,, \tag{11}$$

where $\mathrm{D}[q||f_t] \triangleq \int q(s) \log(q(s)/f_t(s))\,\mathrm{d}s$ is the Kullback-Leibler (KL) divergence (i.e., relative entropy). The KL divergence is an information-theoretic concept that quantifies how much one probability distribution differs from another distribution [40]. By straightforward manipulation, the free energy can be decomposed as follows:

$$\mathcal{F}_t[q] = \underbrace{-\log Z}_{\text{surprise}} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\overline{y}, x, \overline{u})}{p_p(\overline{y}, x, \overline{u}|\underline{y}, \underline{u})}\right]}_{\text{posterior divergence}}, \tag{12}$$

$$= -\log Z + \mathrm{D}[q||p_p], \tag{13}$$

where p_p denoted the exact (Bayesian) posterior, and where $Z = \int f_t(y, x, \overline{u}|\underline{u})\,\mathrm{d}\overline{y}\,\mathrm{d}x\,\mathrm{d}\overline{u}$, with substituted past observations $\underline{y}$ and controls $\underline{u}$; and $p_p(\overline{y}, x, \overline{u}|\underline{y}, \underline{u}) = \frac{1}{Z} f_t(y, x, \overline{u}|\underline{u})$. Since the posterior (KL) divergence term is always positive, the free energy provides an upper bound on the exact (Bayesian) surprise. This decomposition is often employed to justify the use of free energy as a tool for (approximate) inference and model selection [41].

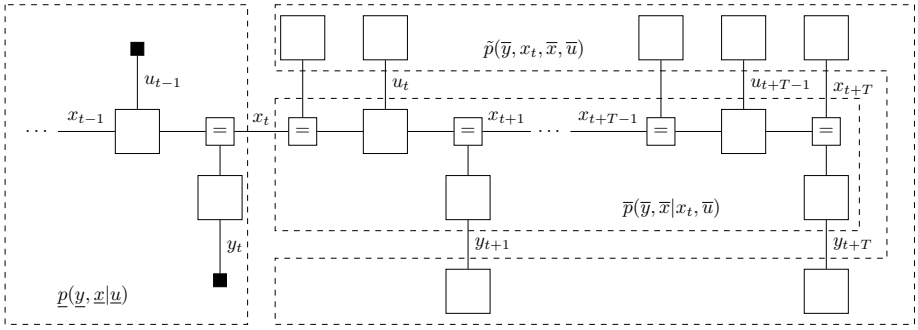

Figure 1: Forney-style factor graph representation of the goal-constrained generative model (9) with indicated factorizations. Observations are indicated by small solid nodes.

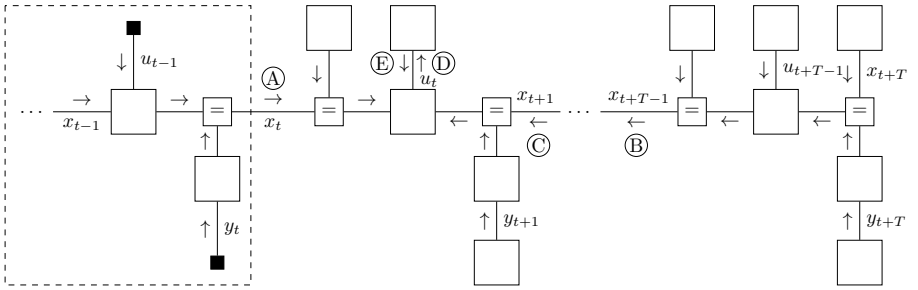

Figure 2: Forney-style factor graph specification of the inference algorithm on the goal-constrained generative model. Here, message Ⓐ represents a state estimate that summarizes past control and observations. The product of messages Ⓓ and Ⓔ yields a posterior belief over the current control, the mode of which is taken as the present action.

Remark 1 (Relation between FEP and surprise)**.** *Generally, the generative model p is a pdf over hidden states (say x) and observations (say y), while q is a pdf only over hidden states. Hence, $p(x, y) = p(y|x)p(x)$, in which $p(x)$ represents a prior. Substituting an observation $y = \boldsymbol{y}$ in the model, yields $f_t(x) = p(y = \boldsymbol{y}|x)p(x)$, which represents the product of a likelihood and the prior. We are interested in obtaining a posterior belief $p_p(x) = f_t(x)/Z$. However, the normalizing constant (Bayesian evidence) $Z = \int f_t(x)\,\mathrm{d}x$ is often intractable to compute, because it involves an integral over all hidden state configurations. As a result, it is often prohibitively expensive (in terms of computational power) to obtain an exact solution for the posterior p_p. Instead, posterior inference is often cast as a free energy optimization problem, where the free energy factorizes as $\mathcal{F}_t[q] = -\log Z + \mathrm{D}[q||p_p]$ as in (12)–(13). Minimization of $\mathcal{F}_t$ thus maximizes Bayesian evidence, while minimizing posterior divergence, making q a close approximation to the (usually intractable) posterior p_p.*

3.3 Control

At time t, the objective is to find the q that minimizes the free energy, i.e.,

$$q_t^* = \arg\min_q \mathcal{F}_t[q]. \tag{14}$$

Looking at (9) and (11), we observe that after optimization, the variational distribution q_t^* simultaneously accounts for the constraints enforced by the system model (2) and the goal prior (10). The posterior for the current control is obtained by marginalization, i.e., $q_t^*(u_t) = \int \cdots \int q_t^*(\overline{y}, x, \overline{u})\,\mathrm{d}\overline{y}, \mathrm{d}x, \mathrm{d}\overline{u}_{\setminus t}$, where $\overline{u}_{\setminus t}$ denotes the sequence obtained by removing u_t from $\overline{u}$. The current action is then chosen as

$$\boldsymbol{u}_{t,q}^* = g(q_t^*)\,, \tag{15}$$

where $g(.)$ is the same as in (5).

3.4 Free Energy Minimization by Message Passing on a Forney-style factor graph

It is instructive to separate inference relating to the past/present from the inference relating to the future. To this end, we substitute the GM $f_t(y, x, \overline{u}|\underline{u})$ of (9) into (11) and use (8c) to factorize the free energy in the following form with separate contributions from the present ($\mathcal{V}_t[q]$) and (expected) future ($\mathcal{G}_t[q]$):

$$\mathcal{F}_t[q] = \underbrace{\mathbb{E}_q\left[\log \frac{q(\underline{x})}{\underline{p}_t(\underline{y} = \underline{\boldsymbol{y}}, \underline{x}|\underline{u} = \underline{\boldsymbol{u}})}\right]}_{\mathcal{V}_t[q]} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\overline{y}, \overline{x}, \overline{u}|\underline{x})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})\,\tilde{p}(\overline{y}, x_t, \overline{x}, \overline{u})}\right]}_{\mathcal{G}_t[q]}. \tag{16}$$

In practice, the optimization of $\mathcal{F}_t$ is often intractable and a specific choice for the factorization of q is made to aid computation [42]. In our case, due to the

factorization of the GM, we can optimize $\mathcal{F}_t$ exactly by belief propagation over the factor graph of the GM [43,44]. To minimize $\mathcal{V}_t$ and $\mathcal{G}_t$, a Forney-style factor graph (FFG) offers a convenient visual representation of a factorized function [45], and is especially well-suited for representing probabilistic models [46]. In an FFG, edges represent variables and nodes (factors) represent relations between variables. The FFG representation of the GM (9) with substituted factorizations (2) and included goal priors (10) is drawn in Fig. 1. The free energy (11) is then minimized by message passing [38, 43, 44, 47] on the FFG representation of the goal-constrained GM. Message passing can be interpreted as first minimizing $\mathcal{V}_t$ and then minimizing a modified version of $\mathcal{G}_t$, based on the outcome of minimizing $\mathcal{V}_t$ [2].

Minimizing $\mathcal{V}_t[q]$

Message passing yields a message Ⓐ (Fig. 2) that represents the current state estimate, given past observations and controls. This message summarizes the information contained within the dashed box:

$$\mu_{\text{Ⓐ}}(x_t) \triangleq \int \underline{p}_t(\underline{y} = \underline{\boldsymbol{y}}, \underline{x} | \underline{u} = \underline{\boldsymbol{u}}) \, \mathrm{d}x_{0:t-1} \,. \tag{17}$$

From the perspective of stochastic optimal control, message Ⓐ connects with the estimator. Moreover, for a linear Gaussian state-space model, the recursive message updates for computing Ⓐ constitute a Kalman filter [48].

Minimizing $\mathcal{G}_t[q]$

In order to minimize $\mathcal{V}_t[q] + \mathcal{G}_t[q]$, we re-normalize the message $\mu_{\text{Ⓐ}}(x_t)$ to obtain a prior $p_e(x_t)$, i.e.,

$$p_e(x_t) \triangleq \frac{1}{C_e} \mu_{\text{Ⓐ}}(x_t) \tag{18}$$

where

$$C_e = \int \underline{p}_t(\underline{y} = \underline{\boldsymbol{y}}, \underline{x} | \underline{u} = \underline{\boldsymbol{u}}) \, \mathrm{d}x_{0:t}. \tag{19}$$

Here, the subscript e emphasizes the fact that p_e represents the pdf of an estimate (of the current state). We then define a modified objective for the expected future

$$\tilde{\mathcal{G}}_t[q] = \mathbb{E}_q \left[\log \frac{q(\overline{y}, \overline{x}, \overline{u}, x_t)}{p_e(x_t) \overline{p}_t(\overline{y}, \overline{x} | x_t, \overline{u}) \, \tilde{p}(\overline{y}, x_t, \overline{x}, \overline{u})} \right], \tag{20}$$

which can again be minimized by message passing. This yields messages Ⓑ – Ⓔ (Fig. 2) by a backward recursion (smoothing pass) over the GM of future variables. The product of Ⓓ and Ⓔ then leads to a marginal belief $q_t^*(u_t)$. Then, the current control action is obtained using (15).

4 From Active Inference To Stochastic Optimal Control

The main question we're interested in is the following: "*When does* (15) *provide the same control actions as* (5)*?*". In other words, when can the ActInf framework be used to solve the traditional stochastic control problem? Below, we investigate this question. Since the goal priors only appear in $\tilde{\mathcal{G}}_t$ and $\mathcal{V}_t$ can be minimized independently, we focus exclusively on the minimization of $\tilde{\mathcal{G}}_t$. We formulate two conditions under which minimizing $\tilde{\mathcal{G}}_t$ reduces to minimizing the stochastic optimal control objective (4).

Note that $\tilde{\mathcal{G}}_t[q]$ can be written as

$$\tilde{\mathcal{G}}_t[q] = \mathbb{E}_q\left[\log \frac{q(\overline{y}, \overline{x}, \overline{u}, x_t)}{p_e(x_t)\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})}\right] - \mathbb{E}_q[\log \tilde{p}(\overline{y}, x_t, \overline{x}, \overline{u})] \ .$$

Then, minimizing $\tilde{\mathcal{G}}_t[q]$ is equivalent to minimizing $\mathcal{G}_t^\dagger[q]$:

$$\mathcal{G}_t^\dagger[q] = \mathbb{E}_q\left[\log \frac{q(\overline{y}, \overline{x}, \overline{u}, x_t)}{p_e(x_t)\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})}\right] + \lambda \mathbb{E}_q\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right] \ , \tag{21}$$

where we substituted the goal prior from (10) and omitted the additive constants that do not depend on the optimization variables.

A striking difference between the optimal control and ActInf objective is that the optimal control objective (4) involves an expectation w.r.t. the system model p_t and policy mapping π, while the free energy involves an expectation w.r.t. the variational distribution q. In order to compare the solutions, we need to create an equal footing.

4.1 Rewriting $\mathcal{G}_t^\dagger$

We start by noting that all arguments of $q(\overline{y}, x, \overline{u})$ that are not within the expectation brackets in (21) are marginalized, i.e., $\underline{x}_{\setminus t}$ is marginalized. Therefore, $\tilde{\mathcal{G}}_t$ is effectively only optimized with respect to $q(\overline{y}, \overline{x}, \overline{u}, x_t)$. We then rewrite the variational distribution in terms of the policy by making use of a region-based approximation [43, 49]. Note that, for a model that is a tree (which is the case for f_t), the region-based approximation is exact. Without loss of generality, we write:

$$\begin{aligned} q(\overline{y}, \overline{x}, \overline{u}, x_t) &= \frac{\prod_{k=t}^{t+T-1} q(y_{k+1}, x_k, x_{k+1}, u_k)}{\prod_{k=t+1}^{t+T-1} q(x_k)} \\ &= \underbrace{\left[\prod_{k=t}^{t+T-1} q(u_k)\right]}_{\overline{\pi}(\overline{u})} \underbrace{\left[\frac{\prod_{k=t}^{t+T-1} q(y_{k+1}, x_k, x_{k+1}|u_k)}{\prod_{k=t}^{t+T-1} q(x_k)}\right]}_{\phi(\overline{y}, \overline{x}|x_t, \overline{u})} q(x_t) \, , \end{aligned} \tag{22}$$

where we simply applied the Bethe factorization to write the variational distribution in terms of a control posterior $\overline{\pi}$, a system posterior ϕ, and the current-state posterior $q(x_t)$.

We now use (22) to rewrite the second term of (21), as:

$$\lambda \mathbb{E}_q \left[\sum_{k=t}^{t+T} \ell(x_k, u_k) \right] = \lambda \mathbb{E}_q \left[\frac{p_e(x_t)}{p_e(x_t)} \frac{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \sum_{k=t}^{t+T} \ell(x_k, u_k) \right] \tag{23a}$$

$$= \lambda \mathbb{E}_{\overline{\pi}} \left[\mathbb{E}_{q(x_t),\phi} \left[\frac{p_e(x_t)}{p_e(x_t)} \frac{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \sum_{k=t}^{t+T} \ell(x_k, u_k) \right] \right], \tag{23b}$$

$$= \lambda \mathbb{E}_{p_e, \overline{p}_t, \overline{\pi}} \left[\frac{q(x_t)}{p_e(x_t)} \frac{\phi(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \sum_{k=t}^{t+T} \ell(x_k, u_k) \right], \tag{23c}$$

where in (23b) we made the expectations due to different terms of $q(\overline{y}, \overline{x}, \overline{u}, x_t) = \overline{\pi}(\overline{u})\phi(\overline{y}, \overline{x}|x_t)q(x_t)$ from (22) explicit and in the last step we interchanged distributions in the expectation subscript with distributions in the numerators, i.e., we used $\mathbb{E}_q\left[\frac{p(s)}{p(s)} f(s)\right] = \mathbb{E}_p\left[\frac{q(s)}{p(s)} f(s)\right]$ for a function f and probability distributions q and p.

We now turn to the first term of (21). Again using the factorization $q(\overline{y}, \overline{x}, \overline{u}, x_t) = \overline{\pi}(\overline{u})\phi(\overline{y}, \overline{x}|x_t)q(x_t)$ from (22), we rewrite this first term as the sum of the negative policy entropy and a posterior divergence:

$$\mathbb{E}_q \left[\log \frac{q(\overline{y}, \overline{x}, \overline{u}, x_t)}{p_e(x_t)\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \right] = \tag{24}$$

$$\mathbb{E}_q[\log \overline{\pi}(\overline{u})] + D(q(x_t) \| p_e(x_t)) + \mathbb{E}_q \left[\log \frac{\phi(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \right].$$

Substituting (23c) and (24) in (21) then reveals the full expression for the ActInf controller objective:

$$\mathcal{G}_t^\dagger[q] = \mathbb{E}_q[\log \overline{\pi}(\overline{u})] + D(q(x_t) \| p_e(x_t)) + \mathbb{E}_q \left[\log \frac{\phi(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \right] +$$

$$\lambda \mathbb{E}_{p_e, \overline{p}, \overline{\pi}} \left[\frac{q(x_t)}{p_e(x_t)} \frac{\phi(\overline{y}, \overline{x}|x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} \sum_{k=t}^{t+T} \ell(x_k, u_k) \right]. \tag{25}$$

Remark 2 (Interpretation of the FEP objective)**.** *The FEP objective can be seen as a trade-off between two terms: one pulls q towards p (under uninformative future values for the controls) and another that minimizes the cost $\sum_{k=t}^{t+T} \ell(x_k, u_k)$. Minimization of only the first three terms in* (25) *(i.e. the case with $\lambda = 0$) leads to an undetermined variational distribution q^*. Namely, because the first three terms in* (25) *directly stem from the first term of Eqn.* (21)*, we have $q^*(\overline{y}, \overline{x}, \overline{u}, x_t) = p_e(x_t)\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})$. Then, we have $\mathcal{G}_t^\dagger[q^*] = 0$ (with $\lambda = 0$).*

Conversely, minimization of only the last term in (25) *(with $\lambda > 0$) leads to a degenerate variational distribution q^* with mass only at a global minimizer*

of $\sum_{k=t}^{t+T} \ell(x_k, u_k)$*, because this last term is equal to* $\mathbb{E}_q\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right]$ *(recall that the last term in* (25) *is only a rewritten version of the last term in* (21)*). In particular, while minimizing* $\mathbb{E}_q\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right]$ *over* q*, since there are no other constraints on* q*, and* q *can be directly chosen as a distribution with point mass at a global minimizer of* $\sum_{k=t}^{t+T} \ell(x_k, u_k)$*. For instance, with* $\ell(.)$ *defined as* (7)*,* q *with point mass at* $x_k = 0$*,* $u_k = 0\ \forall k$ *is a minimizer.*

We now present two sets of conditions under which the actions chosen by the ActInf agent by (15) are the same as the stochastic control actions chosen using (5).

4.2 First Condition: Deterministic Model with Point-Estimate

Let mode$(\cdot)$ represent the mode of a distribution, where ties between multiple modes (if any) are resolved by uniform random selection.

Theorem 1. *Let (i)* $\lambda > 0$*, (ii)* $\phi = \overline{p}_t$*, (iii)* $q_e = p_e$*, and (iv)* $g(\cdot) = \text{mode}(\cdot)$*. Then,* $\boldsymbol{u}^*_{t,q} = \boldsymbol{u}^*_{t,\pi}$*.*

Proof. See Appendix A. □

The below result shows that Theorem 1 implies that the optimal solution is recovered for deterministic systems with a point-estimate.

Corollary 1. *The conditions (ii)* $\phi = \overline{p}_t$*, and (iii)* $q_e = p_e$ *of Theorem 1 occur in the case of a deterministic model* $\overline{p}_t$ *in conjunction with a point estimate for the current state.*

Proof. In the case of a deterministic model, $\overline{p}_t$ of (1) is constrained to delta functions; i.e., $p(x_{k+1}|x_k, u_k) = \delta(x_{k+1} - f_x(x_k, u_k))$ and $p(y_k|x_k) = \delta(y_k - f_y(x_k))$, for some deterministic functions $f_x(\cdot)$ and $f_y(\cdot)$. A point estimate for the current state (after the minimization of $\mathcal{V}_t$) is chosen as $p_e(x_t) \triangleq \delta(x_t - \hat{\boldsymbol{x}}_t)$. Then, any condition other than (ii) $\phi = \overline{p}_t$, (iii) $q_e = p_e$ will lead to infinite divergence in (24), and hence in (25). By contradiction, (ii) $\phi = \overline{p}_t$ and (iii) $q_e = p_e$ are the only viable solutions to the minimization of $\tilde{\mathcal{G}}_t$ under the choice of a deterministic model with a point estimate for the current state. □

4.3 Second Condition: Vanishing State and Control Cost

We now consider minimizers of (25) as a function of λ, and define

$$r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u}) \triangleq \frac{q^*(x_t)\phi^*(\overline{y}, \overline{x}|x_t, \overline{u})}{p_e(x_t)\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})} = \frac{q^*(\overline{y}, \overline{x}, x_t|\overline{u})}{p_t(\overline{y}, \overline{x}, x_t|\overline{u})}. \tag{26}$$

Note that the distribution q is an argument of (25). Hence, the optimal q depends on λ. In light of Remark 2, we see that for $\lambda = 0$, $r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u}) = 1$, while for $\lambda > 0$, $r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u}) \neq 1$.

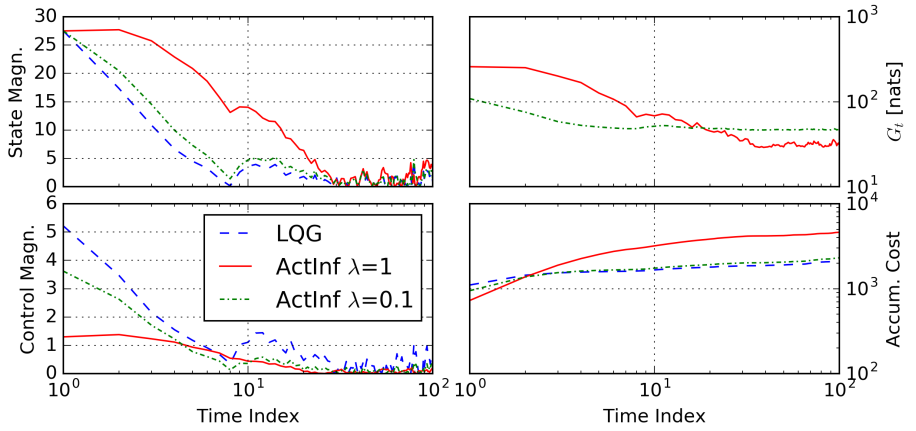

Figure 3: Results comparing LQG with ActInf control, where the time-axis is log-scaled. The more aggressive LQG control (bottom left), leads to faster state adjustments (top left). ActInf control for small but nonzero λ reduces to LQG control. Notably, although ActInf control with $\lambda = 1$ accumulates higher cost in terms of $\ell(x_k, u_k)$ in (7) (bottom right), it achieves lower free energy than ActInf control with small λ (top right).

Theorem 2. *Let (i)* $\lim_{\lambda\to 0^+} \frac{1}{\lambda}\mathbb{E}_q[\log r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u})] = 0$*, (ii)* $\lim_{\lambda\to 0^+} r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u}) = 1$*,* $\forall \overline{y}, \overline{x}, x_t, \overline{u}$*, and (iii)* $\boldsymbol{u}_t = \operatorname{mode} \pi_t$*. Then,* $\boldsymbol{u}^*_{t,q} = \boldsymbol{u}^*_{t,\pi}$*.*

Proof. See Appendix B. □

Condition (ii) requires that, under a vanishing λ, the second term of (25) grows to zero faster than λ itself. Hence, under (ii), the last term will dominate over the second term, retaining the dependence of $\tilde{\mathcal{G}}_t$ on ℓ (see the proof for details). Note that if we outright require $\lambda = 0$, all dependence on ℓ is immediately lost. Instead, the limit ensures that the influence of the cost ℓ is retained.

It is not straightforward to see when the conditions (i) – (ii) of Theorem 2 apply. In the subsequent sections, we further discuss the implications of Theorem 1 and Theorem 2 for the special case of a linear Gaussian SSM.

5 Relationship Between LQG Control And Active Inference For A Linear Gaussian SSM

We now investigate the behavior of the ActInf controller under a linear Gaussian state-space model. We assume a linear Gaussian system with the respective transition and observation precisions W_w and W_v, as follows:

$$p(x_{t+1}|x_t, u_t) = \mathcal{N}(x_{t+1}|Ax_t + Bu_t, W_w^{-1}) \tag{27a}$$

$$p(y_t|x_t) = \mathcal{N}(y_t|Cx_t, W_v^{-1})\,. \tag{27b}$$

where the notation $\mathcal{N}(z|m, V)$ represents a Gaussian distribution with the mean m and the covariance (inverse precision) matrix $V = W^{-1}$ for the variable z. We consider the quadratic cost in (7), leading to independent Gaussian goal priors (10).

5.1 Algebraic Results for the Active Inference Controller

A closed-form expression for the resulting ActInf regulator is obtained by propagating the messages of Fig. 2 algebraically as follows:

Theorem 3. *The ActInf solution for the system in* (27) *is given by*

$$\boldsymbol{u}^*_{t,q} = -K_t \hat{\boldsymbol{x}}_t \tag{28a}$$

$$K_t = \left[B^{\mathrm{T}} \Big(A \hat{V}_t' A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1} \Big)^{-1} B + \lambda R \right]^{-1} \times B^{\mathrm{T}} \Big(A \hat{V}_t' A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1} \Big)^{-1} A \, \hat{V}_t' \, \hat{W}_t \tag{28b}$$

$$\hat{V}_t' = \Big(\hat{W}_t + \lambda Q \Big)^{-1} , \tag{28c}$$

where $\hat{\boldsymbol{x}}_t$ and $\hat{W}_t$ are the respective mean and precision of p_e (the normalized message Ⓐ). Here, P_k, i.e. the precision of the backward message over state x_k, is given by

$$P_{t+T} = \lambda Q \tag{29a}$$

$$P_{k-1} = -A^{\mathrm{T}} P_k B \left(R' + B^{\mathrm{T}} P_k B \right)^{-1} B^{\mathrm{T}} P_k A + A^{\mathrm{T}} P_k A + \lambda Q \tag{29b}$$

$$R' = \Big([\lambda R]^{-1} + \left[B^{\mathrm{T}} W_w B \right]^{-1} \Big)^{-1} . \tag{29c}$$

Proof. See Appendix C. □

Note that this result provides an iterative procedure for finding the ActInf solution. In particular, we initialize P_{t+T} with (29a). Then, P_k's can be calculated iteratively and offline, i.e., without obtaining the measurements. Then, the action is found using (28). Here, calculation of K_t requires calculation of $\hat{W}_t$. In the LQG case, p_e is a Gaussian pdf with a mean and covariance that are given by the standard Kalman filtering equations, see for instance [48, 50].

We now investigate the conditions implied by the two theorems and the corollary from Sec. 3.1.2. Theorem 1 assumes a deterministic model and a point estimate for the current state (Cor. 1), which corresponds to $\hat{W}_t = W_w = \epsilon I_2, \epsilon \to \infty$. Theorem 2 investigates the dependence on λ, and lets $\lambda \to 0^+$. In both cases (29c) reduces to $R' = \lambda R$, and (28b) reduces to $K_t = \left[B^{\mathrm{T}} P_{t+1} B + \lambda R \right]^{-1} B^{\mathrm{T}} P_{t+1} A$, thus recovering the classically optimal LQG solution in the form of the discrete-time finite horizon Ricatti equations [34]. Note that compared to the standard LQG solution, both Q and R appear to be

scaled with λ in the above equations, which has no effect on the optimal solution. This can be seen for instance by recognizing that this scaling corresponds to the scaling of both matrices with λ in (7), which corresponds to a simple scalar scaling of the objective function.

6 Numerical Results

6.1 Scenario

In this section, we illustrate the performance of the ActInf controller for varying positive values of λ and compare the results with the standard LQG scenario. The ActInf simulations are performed with the ForneyLab probabilistic programming toolbox [51], and follow the experimental protocol in [38]. The protocol at each time t consists of three main steps: (i) find Ⓐ and the current state estimate p_e by minimizing $\mathcal{V}_t$ (16), (ii) from the estimate, find a control posterior $q_t^*(u_t)$ by minimizing $\tilde{\mathcal{G}}_t$ (20), and (iii) pass a selected action to the system (1) to obtain a new observation.

For the system, we use (27), with $C = R = Q = W_v = W_w = I_2$, $A = \begin{pmatrix} 1 & 0.1 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 0.1 & 0.5 \\ 0.05 & 0.5 \end{pmatrix}$. The GM follows the system assumptions and uses a lookahead of $T = 10$. We initialize the system relatively far from equilibrium, at $x_0 = (25, 25)^{\mathrm{T}}$ and choose a vague prior for the initial state x_0.

6.2 Discussion

The results are presented in Fig. 3, which leads to several interesting observations. Firstly, for small but nonzero λ, the results (controls, state trajectory, the accumulated cost for $\ell(x_k, u_k)$ and also $\mathcal{G}_t$) of the ActInf controller approaches the results of the LQG controller as expected; see Sec. 3.1.2 and also the discussions at the end of Sec. 5. We note that, for the current system, $\lambda = 0.01$ (not plotted) already renders the results of the ActInf and LQG controller nearly visually indistinguishable.

Secondly, the LQG controller is more aggressive than the ActInf regulator in terms of the controls, i.e., the magnitude of the LQG controls are relatively large compared to those of the ActInf regulator. The explicit inclusion of the process noise in ActInf is in contrast to the LQG scenario where the process and estimation noise only affect the state estimation directly but not the regulator [16]. In particular, (29a–28c) depend explicitly upon W_w and $\hat{W}_t$, whereas these terms are not present in the original Ricatti equations. These terms make the ActInf controller more conservative.

Thirdly, the accumulated cost in terms of $\ell(x_k, u_k)$ for the ActInf controller approaches the optimal cost of the LQG controller under decreasing λ. This observation is consistent with Theorem 2, which formulates sufficient conditions for making the ActInf solution coincide with the LQG solution. Interestingly, and perhaps counter intuitively, the terminal free energy for $\lambda = 1$ is improved

(lower) compared to the $\lambda = 0.1$ case. This effect can be interpreted in light of the *good regulator theorem*, which states that "every good regulator of a system must be a model of that system" [52]. Namely, where the LQG cost function (7) measures a quadratic cost, the free energy (25) offers an approximate measure of model fitness (12). This then implies that the ActInf regulator with $\lambda = 1$ better models the system properties than the ActInf regulator with $\lambda = 0.1$, leading to lower free energy.

7 Conclusions

ActInf and the free energy principle provide a flexible and general framework for stochastic optimal control problems. By including the control cost as goal priors, the control cost appears as an additive term in the free energy. The resulting free energy minimization problem can be solved by belief propagation over the associated factor graph, leading to an elegant and tractable approach to solve stochastic optimal control problems. In general, the ActInf controller does not solve the underlying stochastic optimal control problem. To address this, we provide sufficient conditions for which ActInf reduces to traditional stochastic optimal control. In other words, under certain conditions, stochastic optimal control is a subset of ActInf control. Finally, while it is not known for which classes of problem the sufficient conditions hold, we prove and numerically demonstrate that the ActInf controller is a generalization of the important case of the LQG controller.

At the heart of these methods lies the fact that ActInf allows us to directly control the modeling assumptions. Therefore, we can explicitly include the anticipated effect of the costs and the noise in the control policy. Controlling these assumptions allows us to reproduce traditional stochastic optimal control solutions, such as the LQG controller.

References

[1] H.-A. Loeliger, J. Dauwels, V. M. Koch, and S. Korl, "Signal processing with factor graphs: examples," in *First International Symposium on Control, Communications and Signal Processing, 2004.*, pp. 571–574, IEEE, 2004.

[2] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[3] A. Swami, Q. Zhao, Y. Hong, and L. Tong, *Graphical Models and Fusion in Sensor Networks*, pp. 215–249. John Wiley & Sons, 2007.

[4] G. Ferri, A. Munafò, A. Tesei, P. Braca, F. Meyer, K. Pelekanakis, R. Petroccia, J. Alves, C. Strode, and K. LePage, "Cooperative robotic

networks for underwater surveillance: an overview," *IET Radar, Sonar & Navigation*, vol. 11, no. 12, pp. 1740–1761, 2017.

[5] F. Meyer, H. Wymeersch, M. Fröhle, and F. Hlawatsch, "Distributed estimation with information-seeking control in agent networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2439–2456, 2015.

[6] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.

[8] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 809–819, 2005.

[9] P. E. Caines, *Linear stochastic systems*, vol. 77. SIAM, 2018.

[10] M. Kárnỳ, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.

[11] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," *AAAI*, 2008.

[12] H. J. Kappen, V. Gomez, and M. Opper, "Optimal control as a graphical model inference problem," *Machine Learning*, vol. 87, pp. 159–182, May 2012.

[13] J. Watson, H. Abdulsamad, and J. Peters, "Stochastic Optimal Control as Approximate Input Inference," *Conf. on Robot Learning*, 2019.

[14] S. Levine, "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review," *arXiv:1805.00909*, 2018.

[15] M. Toussaint, "Robot trajectory optimization using approximate inference," in *Int. Conf. on Machine Learning (ICML)*, pp. 1–8, 2009.

[16] C. Hoffmann and P. Rostalski, "Linear Optimal Control on Factor Graphs - a Message Passing Perspective," in *20th IFAC World Congress*, (Toulouse, France), July 2017.

[17] J. H. Lee, "Model predictive control: Review of the three decades of development," *IJCAS*, vol. 9, no. 3, p. 415, 2011.

[18] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[19] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[20] T. Degris, P. M. Pilarski, and R. S. Sutton, "Model-free reinforcement learning with continuous action in practice," in *2012 American Control Conference (ACC)*, pp. 2177–2182, 2012.

[21] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic MPC for model-based reinforcement learning," in *2017 IEEE ICRA*, pp. 1714–1721, May 2017.

[22] S. Kamthe and M. P. Deisenroth, "Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control," *arXiv:1706.06491*, June 2017.

[23] K. J. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology, Paris*, vol. 100, pp. 70–87, Sept. 2006.

[24] M. J. D. Ramstead, P. B. Badcock, and K. J. Friston, "Answering Schrödinger's question: A free-energy formulation," *Physics of Life Reviews*, 2018.

[25] B. Sallans and G. E. Hinton, "Using free energies to represent Q-values in a multiagent reinforcement learning task," in *Adv. in neural information process. systems*, pp. 1075–1081, 2001.

[26] K. J. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, pp. 127–138, Feb. 2010.

[27] G. Oliver, P. Lanillos, and G. Cheng, "Active inference body perception and action for humanoid robots," *Arxiv:1906.03022*, 2019.

[28] O. Çatal, J. Nauta, T. Verbelen, P. Simoens, and B. Dhoedt, "Bayesian policy selection using active inference," *arXiv:1904.08149*, 2019.

[29] K. Ueltzhöffer, "Deep Active Inference," *Biological Cybernetics*, vol. 112, pp. 547–573, Dec. 2018.

[30] S. Schwöbel, S. Kiebel, and D. Markovic, "Active Inference, Belief Propagation, and the Bethe Approximation," *Neural Computation*, vol. 30, pp. 2530–2567, Sept. 2018.

[31] M. Baltieri and C. L. Buckley, "Active Inference: Computational Models of Motor Control without Efference Copy," in *2019 Conf. on Cognitive Computational Neuroscience*, 2019.

[32] B. Millidge, A. Tschantz, A. K. Seth, and C. L. Buckley, "On the relationship between active inference and control as inference," in *1st International Workshop on Active Inference*, 2020.

[33] A. Imohiosen, J. Watson, and J. Peters, "Active inference or control as inference? a unifying view," in *1st International Workshop on Active Inference*, 2020.

[34] T. Glad and L. Ljung, *Control Theory: Multivariable and Nonlinear Methods*. Taylor-Francis, 2000.

[35] H. Attias, "Planning by Probabilistic Inference," in *Inter. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2003.

[36] B. de Vries and K. J. Friston, "A Factor Graph Description of Deep Temporal Active Inference," *Frontiers in Computational Neuroscience*, vol. 11, Oct. 2017.

[37] T. Parr and K. J. Friston, "Generalised free energy and active inference: can the future cause the past?," *bioRxiv*, Apr. 2018.

[38] T. W. van de Laar and B. de Vries, "Simulating Active Inference Processes by Message Passing," *Frontiers in Robotics and AI*, vol. 6, p. 20, 2019.

[39] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, pp. 859–877, Apr. 2017.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[41] H. Attias, "A variational Bayesian framework for graphical models," in *Adv. in neural information process. systems*, pp. 209–215, 2000.

[42] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[43] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, pp. 2282–2312, July 2005.

[44] T. Heskes, "Stable fixed points of loopy belief propagation are local minima of the Bethe free energy," in *Adv. in neural information process. systems*, pp. 359–366, 2003.

[45] G. D. Forney, "Codes on graphs: normal realizations," *IEEE Trans. on Information Theory*, vol. 47, pp. 520–548, Feb. 2001.

[46] H.-A. Loeliger, "An introduction to factor graphs," *Signal Processing Magazine, IEEE*, vol. 21, no. 1, pp. 28–41, 2004.

[47] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE Inter. Symp. on Information Theory*, pp. 2546–2550, June 2007.

[48] S. Korl, *A factor graph approach to signal modelling, system identification and filtering*. ETH Zurich, 2005.

[49] R. Cowell, "Introduction to inference for Bayesian networks," in *Learning in graphical models*, pp. 9–26, Springer, 1998.

[50] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[51] M. Cox, T. W. van de Laar, and B. de Vries, "A factor graph approach to automated design of Bayesian signal processing algorithms," *IJAR*, vol. 104, pp. 185–204, Jan. 2019.

[52] R. C. Conant and W. R. Ashby, "Every good regulator of a system must be a model of that system," *Intl. J. Systems Science*, pp. 89–97, 1970.

Appendix

A Proof of Theorem 1

First, we substitute (ii), (iii) in (25) which removes the second and third term and the factors within the expectation of the last term resulting in

$$\mathcal{G}_t^\dagger[q] = \mathbb{E}_q[\log \overline{\pi}(\overline{u})] + \lambda \mathbb{E}_{p_e, \overline{p}_t, \overline{\pi}}\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right] . \tag{30}$$

Let $L_t^T(\overline{x}, x_t, \overline{u}) \triangleq \sum_{k=t}^{t+T} \ell(x_k, u_k)$. We now focus on the second term in (30)

$$\lambda \int p_e(x_t)\overline{p}(\overline{y}, \overline{x}|x_t, \overline{u})\overline{\pi}(\overline{u})L_t^T(\overline{x}, x_t, \overline{u}) \,\mathrm{d}\overline{u}\,\mathrm{d}\overline{x}\,\mathrm{d}\overline{y}\,\mathrm{d}x_t \tag{31a}$$

$$= C_e\lambda \int \underline{p}_t(\underline{y}, \underline{x}|\underline{u})\overline{p}_t(\overline{y}, \overline{x}|x_t, \overline{u})\overline{\pi}(\overline{u})L_t^T(\overline{x}, x_t, \overline{u}) \,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}\overline{u} \tag{31b}$$

$$= C_e\lambda \int p_t(y, x|u)\overline{\pi}(\overline{u})L_t^T(\overline{x}, x_t, \overline{u}) \,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}\overline{u} \tag{31c}$$

$$= C_e\lambda \mathcal{J}_t \,, \tag{31d}$$

where in (31b), we have used (19) and (18); and in (31c) we have used (8c); and in (31d) we have used (4), i.e. the definition of $\mathcal{J}_t$.

Since the mode of the policy is used for the current action by (iv), the negative policy entropy term $\mathbb{E}_q[\log \overline{\pi}(\overline{u})]$ in (30) does not affect action selection. We therefore absorb the policy entropy in a constant C. Hence, (25) reduces to a function of the form $\mathcal{G}_t^\dagger[q] = \lambda C_e \mathcal{J}_t[q] + C$. Scaling of the scalar optimal control objective $\mathcal{J}_t$ does not affect regulator behavior. Hence, the standard stochastic control solution $\boldsymbol{u}_{t,\pi}^*$ is the same as the ActInf solution $\boldsymbol{u}_{t,q}^*$.

B Proof of Theorem 2

Recall from Theorem 1 that under condition (iii), the first term in (25) does not affect the optimal solution. Furthermore, (ii) removes the ratio $r_\lambda(\overline{y}, \overline{x}, x_t, \overline{u})$

from the last term in (25). Hence, substituting in these modifications and multiplying the objective with $1/\lambda$ (note that multiplications with $1/\lambda > 0$ do not change optimal solutions), we have following objective function:

$$\frac{1}{\lambda} D(q^*(x_t) \| p_e(x_t)) + \frac{1}{\lambda} \mathbb{E}_{q^*}\left[\log \frac{\phi^*(\overline{y}, \overline{x} | x_t, \overline{u})}{\overline{p}_t(\overline{y}, \overline{x} | x_t, \overline{u})}\right] + \mathbb{E}_{p_e^*, \overline{p}^*, \overline{\pi}^*}\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right] \tag{32a}$$

$$= \frac{1}{\lambda} \mathbb{E}_{q^*}\left[\log \frac{q^*(x_t)\phi^*(\overline{y}, \overline{x} | x_t, \overline{u})}{p_e(x_t)\overline{p}_t(\overline{y}, \overline{x} | x_t, \overline{u})}\right] + \mathbb{E}_{p_e^*, \overline{p}^*, \overline{\pi}^*}\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right], \tag{32b}$$

where in (32b) we have used

$$D(q^*(x_t) \| p_e(x_t)) = \mathbb{E}_{q^*(x_t)}\left[\log \frac{q^*(x_t)}{p_e(x_t)}\right] = \mathbb{E}_{q^*}\left[\log \frac{q^*(x_t)}{p_e(x_t)}\right].$$

Taking the limit with $\lambda \to 0^+$ and substituting (ii), we obtain $\mathbb{E}_{p_e^*, \overline{p}^*, \overline{\pi}^*}\left[\sum_{k=t}^{t+T} \ell(x_k, u_k)\right]$ as desired. Then, the optimal control objective (4) is again (proportionally) recovered, and hence $\boldsymbol{u}^*_{t,q} = \boldsymbol{u}^*_{t,\pi}$.

C Proof of Theorem 3

The algebraic result for the ActInf regulator, (29) and (28), is obtained by message passing (Fig. 2). We derive this result by following the standard belief propagation update rules as summarized by [48, Table 4.1]. For notational convenience, we write mean-variance and mean-precision parameterized Gaussian distributions as $\mathcal{N}_V$ and $\mathcal{N}_W$ respectively, where distribution variable arguments are left implicit.

Backward Recursion

The backward recursion (29) follows from message passing in a section of the model as visualized in Fig. 4. Note that our specific choice of goal prior (10) is independent of y_k. As a result, messages ② and ③ are uninformative, and do not contribute to the end result.

The messages of Fig. 4 are computed as follows:

$$\begin{aligned}
&① \propto \mathcal{N}_W(0, P_k) \\
&② \propto 1 \\
&③ \propto 1 \\
&④ \propto \mathcal{N}_W(0, P_k) \\
&⑤ \propto \mathcal{N}_V\left(0, P_k^{-1} + W_w^{-1}\right) \\
&⑥ \propto \mathcal{N}_W(0, \lambda R) \\
&⑦ \propto \mathcal{N}_V\left(0, B(\lambda R)^{-1} B^{\mathrm{T}}\right)
\end{aligned}$$

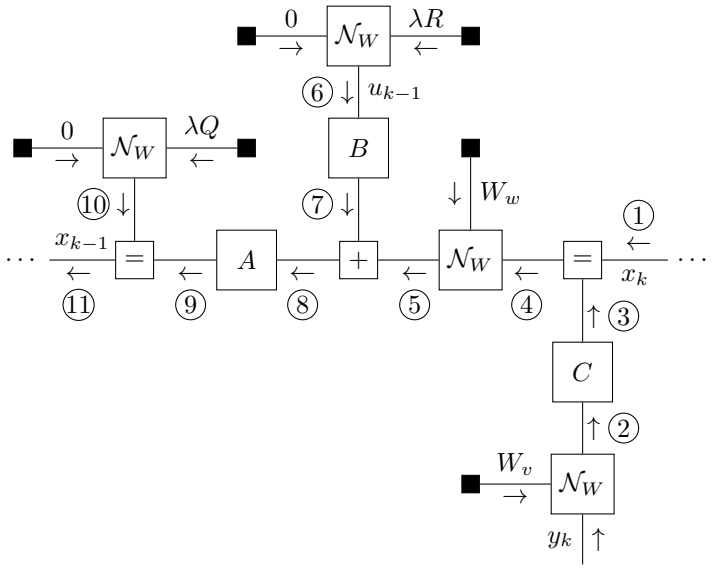

Figure 4: Message passing schedule for recursive backward propagation in a single (future) section of a linear Gaussian state-space model (27).

$$
\begin{aligned}
&⑧ \propto \mathcal{N}_V\big(0, P_k^{-1} + B(\lambda R)^{-1}B^{\mathrm{T}} + W_w^{-1}\big)\\
&⑨ \propto \mathcal{N}_W\big(0, A^{\mathrm{T}}(P_k^{-1} + B(\lambda R)^{-1}B^{\mathrm{T}} + W_w^{-1})A\big)\\
&⑩ \propto \mathcal{N}_W(0, \lambda Q)\\
&⑪ \propto \mathcal{N}_W\big(0, \underbrace{A^{\mathrm{T}}(P_k^{-1} + B(\lambda R)^{-1}B^{\mathrm{T}} + W_w^{-1})A + \lambda Q}_{P_{k-1}}\big)\,,
\end{aligned}
$$

where we identify a recursion over P_{k-1}. The recursion is initialized with $P_{t+T} = \lambda Q$, and terminates when P_{t+1} is computed. The result simplifies further:

$$
\begin{aligned}
P_{k-1} &= A^{\mathrm{T}}(P_k^{-1} + B\overbrace{[(\lambda R)^{-1} + (B^{\mathrm{T}}W_w B)^{-1}]}^{[R']^{-1}}B^{\mathrm{T}})A\\
&\quad + \lambda Q \qquad \textit{(using } BB^{-1} = I\textit{)}\\
&= A^{\mathrm{T}}P_k A - A^{\mathrm{T}}P_k B[R' + B^{\mathrm{T}}P_k B]^{-1}B^{\mathrm{T}}P_k A\\
&\quad + \lambda Q \qquad \textit{(using Woodbury identity)}\,,
\end{aligned}
$$

with

$$
R' = [(\lambda R)^{-1} + (B^{\mathrm{T}}W_w B)^{-1}]^{-1}\,.
$$

This concludes the derivation of (29). The computation of R' can be simplified by making use of Searle's identity.

Control Law

The control law (28) follows from message passing in a section of the model as visualized in Fig. 5.

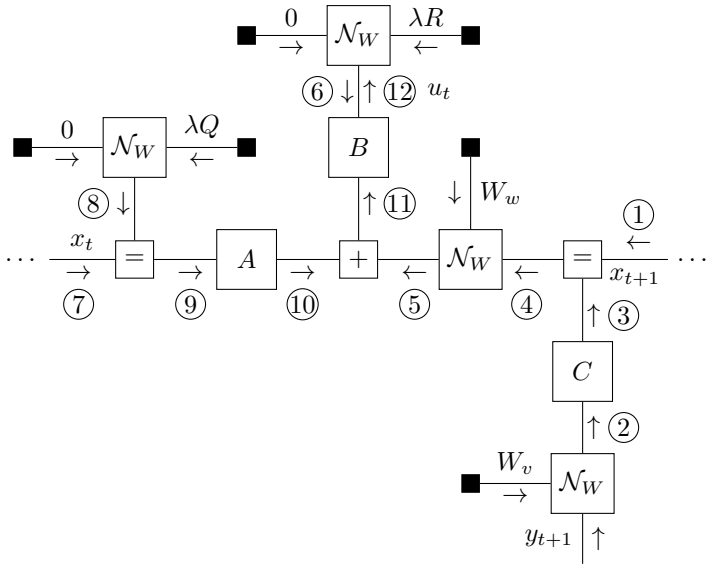

Figure 5: Message passing schedule for the control law in a single (present) section of a linear Gaussian state-space model (27).

The messages of Fig. 5 are computed as follows:

$$
\begin{aligned}
① &\propto \mathcal{N}_W(0, P_{t+1}) \\
② &\propto 1 \\
③ &\propto 1 \\
④ &\propto \mathcal{N}_W(0, P_{t+1}) \\
⑤ &\propto \mathcal{N}_V\big(0, P_{t+1}^{-1} + W_w^{-1}\big) \\
⑥ &\propto \mathcal{N}_W(0, \lambda R) \\
⑦ &\propto \mathcal{N}_W\Big(\hat{\boldsymbol{x}}_t, \hat{W}_t\Big) \\
⑧ &\propto \mathcal{N}_W(0, \lambda Q) \\
⑨ &\propto \mathcal{N}_V\Big([\hat{W}_t + \lambda Q]^{-1}\hat{W}_t\hat{\boldsymbol{x}}_t, [\hat{W}_t + \lambda Q]^{-1}\Big) \\
⑩ &\propto \mathcal{N}_V\Big(A[\hat{W}_t + \lambda Q]^{-1}\hat{W}_t\hat{\boldsymbol{x}}_t, A[\hat{W}_t + \lambda Q]^{-1}A^{\mathrm{T}}\Big) \\
⑪ &\propto \mathcal{N}_V\Big(- A[\hat{W}_t + \lambda Q]^{-1}\hat{W}_t\hat{\boldsymbol{x}}_t, \\
&\qquad A[\hat{W}_t + \lambda Q]^{-1}A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1}\Big)
\end{aligned}
$$

$$\textcircled{12} \propto \mathcal{N}_W\Big(- B^{-1}A[\hat{W}_t + \lambda Q]^{-1}\hat{W}_t\hat{\boldsymbol{x}}_t,$$
$$B^{\mathrm{T}}[A[\hat{W}_t + \lambda Q]^{-1}A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1}]^{-1}B\Big)\,.$$

The current control then follows from

$$\begin{aligned} q_t^*(u_t) &\propto \textcircled{6} \times \textcircled{12} \\ \boldsymbol{u}_t &= \operatorname{mode} q_t^*(u_t) \\ &= -K_t\hat{\boldsymbol{x}}_t\,, \end{aligned}$$

where (using the Gaussian equality rule)

$$\begin{aligned} K_t = [B^{\mathrm{T}}(A\hat{V}_t'A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1})^{-1}B + \lambda R]^{-1}\times \\ B^{\mathrm{T}}(A\hat{V}_t'A^{\mathrm{T}} + P_{t+1}^{-1} + W_w^{-1})^{-1}A\,\hat{V}_t'\,\hat{W}_t\,, \end{aligned}$$

with

$$\hat{V}_t' = (\hat{W}_t + \lambda Q)^{-1}\,.$$

This concludes the derivation of (28).