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

A Probabilistic Treatment of (PO)MDPs with Multiplicative Reward Structure

Tom Lefebvre

Abstract—**The objective of this research is to identify optimal control formulations or similar problems, that can be solved by practising inference on probabilistic graph models instead of solving temporal nonlinear optimization problems. This is an active research topic in the Reinforcement Learning and control community that is better known as Control as Inference. Inference on probabilistic graph models is a computational process that is easily automated for example using message passing. In this contribution we show that Partially Observable Markov Decision Problems with multiplicative reward structure can be represented by an equivalent Maximum Likelihood Estimation problem. Subsequently the estimation problem can be treated by means of the Expectation-Maximization algorithm. We show that maximization of the Evidence Lower Bound can be reinterpreted as a probabilistic control problem which is itself a density matching interpretation of Control as Inference. The associated probabilistic policy can be represented as a conditional density and can be calculated by message passing on the probabilistic graph model. These results provide a unified account of probabilistic control and control as inference with multiplicative reward structures under partial observability.**

I. Introduction

The Reinforcement Learning (RL) and control community at large is occupied with the problem of automated decision making. Therefore it relies on the computational framework of Markov Decision Processes (MDPs). Formally, MDPs are defined by a probabilistic graph model (PGM), describing the underlying dynamical system, augmented with an external notion of reward that quantifies the value of every behavioural pattern described by the PGM. The solution of an MDP is given by an optimal decision-making strategy or so called policy. An agent – tasked with controlling the system or navigating the environment – wielding the optimal policy, is expected to execute its task with maximum reward.

To circumvent the computational challenges associated with solving MDPs, there have been various attempts to encode the notion of reward directly into the PGM used to describe the dynamical model. That way, optimal decision making could be formalized as an inference problem on the extended PGM rather than as a temporal nonlinear optimization problem. These endeavours lead to a research program that we may refer to as *probabilistic control* [1, 2] or, the more common, *Control as Inference* (CaI) [3].

One interpretation of CaI is to encode value in the PGM through an auxiliary set of exogenous binary observation variables. The variables their (future) values are assumed to be known and true and *indicate* that an optimal decision has been made. Thence, the control system is inferred by calculating the probability of making a decision at present time assuming (future) optimally has been achieved.

By a specific choice of the auxiliary emission model, the framework resumes close analogies with the theory of optimal control [4, 5, 6]. Recent work established an explicit connection between CaI and MDPs through the lense of probabilistic control [6]. Probabilistic control formulates CaI as a density matching problem. Depending on the measure used to quantify the density matching condition, a connection exists with MDPs with an additive or a multiplicative reward structure. Furthermore, for MDPs with a multiplicative reward structure, there exist an equivalent Maximum Likelihood estimation (MLE) problem [6, 7]. The associated Evidence Lower Bound (ELBO) corresponds with the probabilistic control problem with multiplicative reward. Finally, it can be shown that the associated probabilistic control policy can be calculated by conditioning the present action on the state and the known future auxiliary observation variables [6].

To the best of our knowledge, there are no references that address CaI or probabilistic control under the restriction of partial observability with the exception of a preliminary result given in [8]. We assume this to be a result of the technicality of the subject and ensuing practical challenges. Such a treatment is however highly desirable since the POMDP framework treats a variety of real-world applications for which the MDP framework is simply inadequate.

In the present paper we establish such a treatment by extending the results from [6]. The present study is limited to (PO)MDPs with multiplicative reward structure or so called risk-sensitive optimal control [9, 10]. Partially because the result of [8] stands for POMDPs with additive reward but mostly because the multiplicative setting alone renders policies that can be evaluated by inference on a PGM.

The ambition of this paper is then simply to provide an overview of the general theory and ideas for both MDPs as well as POMDPs and to highlight connections with the existing body of work. As such we provide a unified account of probabilistic control and CaI associated to multiplicative reward structure that extends to a partially observable setting.

II. Related work

The historical development leading to our present understanding of probabilistic control and CaI has been lengthy and interesting with contributions coming from various research communities. Among the first works to uncover a connection between optimal control and inference on PGMs were given by Kappen and Todorov [11, 12]. Both authors identified a class of stochastic optimal control problems where inference emerges naturally. Meaning that the optimal policy could be calculated by evaluating the expected value of an exponential cost-to-go where the expectation was defined under some prior policy. Toussaint, amongst others, explored the idea to encode the notion of value directly into the PGM using an auxiliary set of binary optimality variables and to associate

T. Lefebvre is with the Dynamic Design Lab (D^{2}LAB) of the Department of Electromechanical, Systems and Metal Engineering, Ghent University, Ghent, Belgium. E-mail: tom.lefebvre@ugent.be.

T. Lefebvre is a member of the core-lab MIRO, Flanders Make, Belgium.

the conditional trajectory with an optimal trajectory [4, 13, 14, 15]. Taking the work of Toussaint, Rawlik et al. proposed a density matching approach, minimizing the relative entropy between the closed-loop model and the extended joint model [5]. This naturally lead to a treatment of MDPs with additive reward structure augmented with an entropy regularization term. Interestingly, a similar or at least closely related idea to the density matching approach of Rawlik, was described much earlier by Kárný [1, 2], coined *probabilistic control*.

Several ideas from CaI were rediscovered by the RL community and developed into maximum entropy RL [16, 17] and Maximum a Posteriori RL [18]. Other studies attempted to leverage the theory to establish sample based trajectory optimizers with application in MPC [19, 20, 21]. Several of these works were motivated by the silent presumption that CaI has an efficacy to incite purposeful exploration, explaining why it is of great interest to learning paradigms such as RL. Such has been empirically verified, though thus far no explicit relation with the underlying MDP was given.

As noted in the introduction, it was shown recently that probabilistic control problems *majorize* MDPs with an additive and multiplicative reward structure [6]. It is implied that we can maintain a sequence of probabilistic policies whose stationary point coincides with the optimal policy. This result establishes an explicit connection in the MDP setting. A first extension to POMDPs was attempted in [8]. A non-veridical prescriptive model decomposition was wielded, intending to render the Bayesian belief a sufficient statistic [10]. Although practical by design, the result is not exact.

In conclusion we mention another interesting connection with the framework of Active Inference (AIF) which was independently developed. AIF is an emerging brain theory in theoretical neuroscience which proposes a unified account of perception, action and learning by the brain [22, 23, 24]. AIF hypothesises that agents maintain a prescriptive model of their environment and act to reduce the (expected) free energy, which roughly means they minimize their subjective experience of surprise. The framework is supported by empirical evidence and has a degree of biological plausibility. An earlier comparison between CaI and AIF pointed out the main similarities and some differences [25]. Any belief held by the AIF agent is modelled by a variational density. The agent can thus choose a variational density that leads to simplified calculations, often a mean-field approximation. More fundamentally, there is no explicit encoding of the notion of reward. Rather, reward is encoded through a non-veridical prescriptive model that is biased towards desired observations. Agents act to align their sensory inputs with biased predictions. As a result, exploration in the context of AIF is said to be goal directed, maximizing an expected information gain. Note that this also implies that the observation model is *hijacked*, rendering the perception ambiguous. More recent treatments of the theory, especially in the context of AIF with predictive horizon, have proposed alternative ways to encode value that align more closely with CaI and probabilistic control [26, 27]. Further research is required to fully understand their relationship. Extending CaI to a partially observable setting will proof useful to that end.

Finally, one could argue, in the spirit of Bayesian brain theories such as AIF, that the prescriptive model in [8] is the variational model wielded by the brain. Though, from a control perspective, such a justification is unsatisfactory.

III. Preliminaries

A. Notation

We refer to the leading or trailing part of a time dependent process with $\underline{x}_t = \{x_0, \dots, x_t\}$, and, $\overline{x}_t = \{x_t, \dots, x_T\}$. The index, t, refers to the final or initial time instance of the corresponding subsequence. We silently assume that a complete sequence starts at time $t = 0$ and ends at time $t = T$ except for control sequences ($\underline{u}_{T-1}$, see later) that start at time $t = 0$ but end at time $t = T - 1$. Throughout we refer to the set of all feasible probability density functions with $\mathcal{P}$. We will rely on the context to imply the arguments and properties of the corresponding function class.

B. Agent, prescriptive model and information pattern

Adopting terminology from RL and AIF, we refer to the entity tasked with the decision process, ergo with making policy, as the agent. The agent may materialize as an embedded controller or as a biological controller such as the brain. We assume that the agent make sense of its environment by means of a veridical prescriptive model. In particular here we adopt a controlled Hidden Markov Model (HMM) (see Fig. 1). The process x_t represents the state of the system and cannot be observed. The process y_t is measured and is called the observation process. The agent can act on its environment by determining and applying controls or decisions, u_t.

The information we grant the agent access to determine a control, is referred to as the information pattern, w_t. We consider the following information patterns.

- The agent receives the observation, y_t, and has access to a memory that stores the system's observable history, $\underline{y}_{t-1}$ and $\underline{u}_{t-1}$. This information pattern is also referred to as partially observable.
- When $w_t = x_t$, the information pattern is referred to as fully observable.

We further assume that the agent has access to the following prescriptive model components

- the initial state, x_0, has density $p(x_0)$
- the transition dynamic satisfies the Markov property and is modelled as $p(x_{t+1}|x_t, u_t)$
- the observation satisfies the Markov property, is independent of the control and is modelled as $p(y_t|x_t)$
- in general the agent's policy may be uncertain and is modelled as the feedback density, $\pi_t(u_t|w_t)$

These allow the agent to express the following joint model.

$$p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T) = p(x_0)p(y_0|x_0) \\ \times \prod\nolimits_{t=0}^{T-1} p(x_{t+1}|x_t, u_t)\pi_t(u_t|w_t)p(y_{t+1}|x_{t+1}) \tag{1}$$

We emphasize that the joint model is parametrised by the policy sequence, $\overline{\pi}_{T-1}$, using subscript, π. Therefore we also refer to this joint model as closed-loop since the control process is governed by a feedback policy sequence.

C. Utility theory and decision criteria

According to the principles of normative decision theory and expected utility theory [28], a rational agent makes policy based on the maximization of the expectation of some measure of reward (or the minimisation of the expectation of some measure of cost). There exist two reward measures that produce a rich mathematical theory.

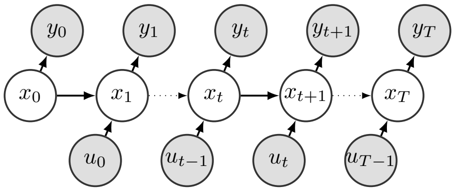

Fig. 1: Probabilistic graph model of a Controlled Hidden Markov Chain. White nodes are observable, grey nodes are hidden.

First we have the classical additive reward, A

$$A = \underline{c}_T = \sum\nolimits_{t=0}^{T-1} c_t(x_t, u_t) + c_T(x_T) \tag{2}$$

Alternatively one can consider the multiplicative measure of reward, M, which is defined as the exponential of the additive reward, A. This is the standard risk-sensitive control objective [9, 10, 29, 30]. In terms of utility theory this means that higher rewards are more valuable to the agent than A leads to suspect. The subjective value of an objective reward measure can be quantified by means of a utility function [31]. If the utility is concave then the agent always prefers a certain return. The agent is said to be risk averse. When the utility is convex, the opposite is true, and the agent attributes disproportional value to higher rewards. The agent is said to be risk seeking. The exponential utility has the advantage that an additive structures translates to a multiplicative one.

$$M = \exp(A) = \exp(\underline{c}_T) \tag{3}$$

D. Probabilistic Control

The probabilistic control framework proposes that an agent acts to find feedback control policy, $\underline{\pi}_{T-1}$, so that the joint closed-loop density model, p_π, is as close as possible to some desired density, p^*. This simple concept can be broken down into two critical questions that must be addressed in order to develop it into a quantifiable and useful theory.

- How does one define a productive desired density?
- How does one define the proximity of densities?

The answer to the first question will be treated in the following two sections. To answer the second question we rely on information projection theory. As proposed by [1, 2, 5], one can use the information-projection where $\mathbb{D}\left[\cdot \parallel \cdot\right]$ denotes the relative entropy or KL-divergence. Noting that the KL-divergence can be expressed as the expectation of the logarithm of the ratio between the modelled and desired joint density, this approach is closely related to utility theory.

$$\min_{\pi \in \mathcal{P}} \mathbb{D}\left[p_\pi \parallel p^*\right] \tag{4}$$

Proximity can also be expressed by means of the reciprocal moment-projection as was pointed out by [6]. This second proximity measure will enjoy our interest henceforth in the context of the multiplicative control objective, M. For a detailed comparison between either proximity measure in the MDP setting, we refer the reader to [6].

$$\min_{\pi \in \mathcal{P}} \mathbb{D}\left[p^* \parallel p_\pi\right] \tag{5}$$

Next we set as our goal to establish an explicit connection between probabilistic control and optimal control theory. We further like to emphasize that results in probabilistic control have been restricted to a fully observable setting.

IV. Probabilistic treatment of MDPs

Utility theory states that a rational agent's makes policy by solving the following optimal control problem. The problem combines the agent's prescriptive model, p_π, with its (subjective) multiplicative reward measure, M. Depending on the information pattern the agent is granted access to this problem corresponds with either an MDP (full observability) or POMDP (partial observability) with multiplicative cost or so called risk-sensitive MDPs or POMDPs.

$$\max_{\underline{\pi}_{T-1} \in \mathcal{P}} \mathbb{E}_{p_\pi}\left[M\right] \tag{6}$$

Starting from the problem above, in this section and the following, we develop a probabilistic treatment of MDPs and POMDPs with multiplicative reward. Although we categorize our results under the probabilistic control framework, arguing that the probabilistic control interpretation is more fundamental, we first establish a connection with Bayesian estimation.

In this section, our treatment is restricted to MDPs. Therefore we can neglect the measurement process, $\underline{y}_T$, altogether. In the next section, results are generalised to POMDPs.

A. Equivalent Bayesian estimation problem

First let us show that the problem in (6) is equivalent to a Maximum Likelihood Estimation (MLE) problem.

To that end we may introduce an auxiliary set of fictitious binary *optimality* variables, $\underline{z}_T$, to encode the external notion of reward [3, 4, 5]. The optimality variables manifest as the measurement variables in the controlled HMM (Fig. 1). Further recall that under the assumption of full observability, the state is also observed. The resulting graph model is given in Fig. 2. For brevity we note z_t when we mean $z_t = \texttt{true}$. To establish a connection with optimal control the following emission model is proposed when the variables, z_t, are `true`

$$p(z_t|x_t, u_t) = \exp(c_t(x_t, u_t)) \tag{7}$$

with the exception of $p(z_T|x_T) = \exp(c_T(x_T))$.

The reader may now verify that, assuming all optimality variables have adopted the given values $\underline{z}_T$, problem (6) and the following MLE problem are equivalent

$$\max_{\underline{\pi}_{T-1} \in \mathcal{P}} p_\pi(\underline{z}_T) \tag{8}$$

where

$$p_\pi(\underline{z}_T) = \mathbb{E}_{p_\pi}\left[p(\underline{z}_T|\underline{x}_T, \underline{u}_{T-1})\right] \tag{9}$$

In the MDP setting, this result was established independently in the following works [6, 7].

B. Expectation-Maximization of the MLE

MLE problems such as problem (8) can be treated by means of the Expectation-Maximization (EM) algorithm. For a detailed exhibition we refer to appendix A. The EM algorithm treats MLE problems recursively. Instead of solving the problem directly, a sequence of approximate problems is solved instead. The approximate problems require maximisation of the Evidence Lower Bound (ELBO). In the present setting the ELBO is given by

$$\begin{aligned} \arg \max_{\underline{\pi}_{T-1} \in \mathcal{P}} & \int p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T) \\ & \times \log p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{z}_T) \mathrm{d}\underline{x}_T \mathrm{d}\underline{u}_{T-1} \end{aligned} \tag{10}$$

where p_ρ refers to the closed-loop joint density parametrised by some prior policy sequence, $\underline{\rho}_{T-1}$.

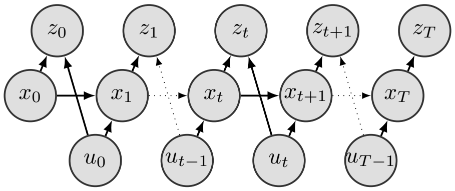

Fig. 2: Probabilistic graph model of extended Controlled Markov Chain. White nodes are observable, grey nodes are hidden.

The solution of problem (10) is given by the following conditional probability. This has been shown by [6].

$$\pi_t(u_t|x_t) = \frac{p_\rho(x_t, u_t|\underline{z}_T)}{p_\rho(x_t|\underline{z}_T)} = p_\rho(u_t|x_t, \underline{z}_T) \tag{11}$$

We emphasize that (11) does not solve (6) but (10), which is a lower bound on (6) by construction. However, we also note that (10) singles out a step from the EM algorithm. This means we can solve problem (8) substituting $\underline{\pi}_{T-1}$ for $\underline{\rho}_{T-1}$. This will result into a new policy sequence which can be substituted again and so forth. By merit of the EM algorithm, this procedure will converge to the deterministic MDP policy.

C. Equivalent probabilistic control problem

At this point, we can address the open question from section III-D. To that end, we first acknowledge that problem (10) can be reinterpreted as the minimization of the moment projection between the posterior, $p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T)$, and the joint model, $p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{z}_T)$. Second, we note that the resulting problem is equivalent to the moment-projection of the posterior, $p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T)$, and, the closed-loop density $p_\pi(\underline{x}_T, \underline{u}_{T-1})$. It follows that

$$\begin{aligned}(10) &\equiv \arg\min_{\underline{\pi}_T \in \mathcal{P}} \mathbb{D}\left[p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T) \parallel p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{z}_T)\right] \\ &\equiv \arg\min_{\underline{\pi}_{T-1} \in \mathcal{P}} \mathbb{D}\left[p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T) \parallel p_\pi(\underline{x}_T, \underline{u}_{T-1})\right]\end{aligned} \tag{12}$$

The second problem can be interpreted as a probabilistic control problem with desired closed-loop density

$$p^*(\underline{x}_T, \underline{u}_{T-1}) = p_\rho(\underline{x}_T, \underline{u}_{T-1}|\underline{z}_T = \underline{\texttt{true}}_T) \tag{13}$$

This observation establishes an equivalence between various probabilistic treatments of optimal control theory.

D. Dynamic programming and message passing on graphs

Finally, we note that (11) can be evaluated using dynamic programming. We refer to [6] for further details. In the next section we will explicitly detail the procedure for POMDPs.

$$\pi_t(u_t|x_t) = \rho_t(u_t|x_t)\frac{p_\rho(\overline{z}_t|x_t, u_t)}{p_\rho(\overline{z}_t|x_t)} \tag{14}$$

The densities $p_\rho(\overline{z}_t|x_t, u_t)$ and $p_\rho(\overline{z}_t|x_t)$ are governed by a backward message passing procedure. Let us define

$$\begin{aligned}Q_t^\bullet(x_t, u_t) &= p_\rho(\overline{z}_t|x_t, u_t) \\ V_t^\bullet(x_t) &= p_\rho(\overline{z}_t|x_t)\end{aligned} \tag{15}$$

Then it can be shown that [6]

$$\begin{aligned}Q_t^\bullet(x_t, u_t) &= p(z_t|x_t, u_t) \\ &\times \int p(x_{t+1}|x_t, u_t)V_{t+1}^\bullet(x_{t+1})\mathrm{d}x_{t+1}\end{aligned} \tag{16}$$

and

$$V_t^\bullet(x_t) = \int \rho_t(u_t|x_t)Q_t^\bullet(x_t, u_t)\mathrm{d}u_t \tag{17}$$

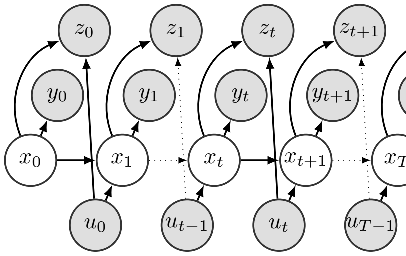

Fig. 3: Probabilistic graph model of extended Controlled Hidden Markov Chain under veridical probability measure, p_π. White nodes are observable, grey nodes are hidden.

V. Probabilistic treatment of POMDPs

The goal of this section is to reiterate the probabilistic treatment of problem (6) under the restriction of partial observability, ergo a probabilistic treatment of POMDPs.

To that end we will make use of the concept of the reference measure, see [10] and references therein. The use of a reference measure involves changing the probability density from p_π to q_π so that the measurement process is decoupled from the state process. This is accounted for by changing the reward measure. It is easily verified that the following optimal control problem is equivalent to (6) (again see [10])

$$\max_{\underline{\pi}_{T-1} \in \mathcal{P}} \mathbb{E}_{q_\pi}\left[\exp(\underline{c}_T + \underline{l}_T)\right] \tag{18}$$

where

$$\begin{aligned}q_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T) &= p(x_0)q(y_0) \\ &\times \prod_{t=0}^{T-1} p(x_{t+1}|x_t, u_t)\pi_t(u_t|w_t)q(y_{t+1})\end{aligned} \tag{19}$$

and

$$\underline{l}_T = \log\frac{p_\pi}{q_\pi} \tag{20}$$

so that

$$l_t(x_t, y_t) = \log\frac{p(y_t|x_t)}{q(y_t)} \tag{21}$$

We emphasize the use of a reference measure by changing the expression for any density affected by it from p to q.

A. Equivalent Bayesian estimation problem

Again, first we show that problem (18) is equivalent to a Maximum Likelihood Estimation (MLE) problem. Here also we introduce an auxiliary set of fictitious binary optimality variables, $\underline{z}_T$, with

$$q(z_t|x_t, u_t, y_t) = \exp(c_t(x_t, u_t) + l_t(x_t, y_t)) \tag{22}$$

except for $q(z_T|x_T, y_T) = \exp(c_T(x_T) + l_T(x_T, y_T))$.Under the veridical probability measure, p_π, this results into the graph model from Fig. 3. The corresponding graph model under the reference measure, q_π, is given in Fig. 4.

The reader may then verify that the following MLE is indeed equivalent to problem (18)

$$\max_{\underline{\pi}_{T-1} \in \mathcal{P}} q_\pi(\underline{z}_T) \tag{23}$$

where

$$q_\pi(\underline{z}_T) = \mathbb{E}_{q_\pi}[q(\underline{z}_T|\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T)] \tag{24}$$

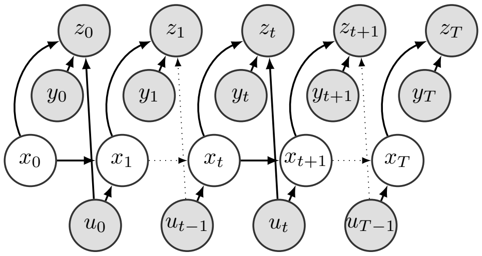

Fig. 4: Probabilistic graph model of extended Controlled Hidden Markov Chain under reference probability measure, q_π. White nodes are observable, grey nodes are hidden.

B. Expectation-Maximization of the MLE

Again let us treat the MLE in (23) with the EM algorithm. This generates the following ELBO

$$\begin{aligned}\arg \max_{\underline{\pi}_{T-1}\in\mathcal{P}} \int & q_\rho(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T | \underline{z}_T) \\ & \times \log q_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T, \underline{z}_T) \mathrm{d}\underline{x}_T \mathrm{d}\underline{u}_{T-1} \mathrm{d}\underline{y}_T\end{aligned} \tag{25}$$

The solution of this problem is given by the following conditional probability. This is easily verified noting that the expectation in (25) decomposes in T separate optimization problems. Then, since by construction, π_t depends on u_t and w_t all other variables are marginalized out. Finally, taking the normalization condition into account one verifies that

$$\begin{aligned}\pi_t(u_t|w_t) &= \frac{q_\rho(\underline{u}_t, \underline{y}_t | \underline{z}_T)}{q_\rho(\underline{u}_{t-1}, \underline{y}_t | \underline{z}_T)} \\ &= q_\rho(u_t|w_t, \underline{z}_T)\end{aligned} \tag{26}$$

This result is as simple as it is intriguing. It directly follows that (26) can be evaluated by applying inference on the graph model in Fig. 4. Though it will be shown (sec. V-D) that the resulting message passing procedure is of higher complexity than the message passing procedure in sec. IV-D. This is a result of the structural complexity of the graph, in Fig. 4.

Further note that the same remarks apply to the policy in (26) and the MLE or POMDP problems in (23) or (18), as were given in section IV-B for MDPs. This observation implies that application of the EM procedure will eventually produce the deterministic POMDP policy.

C. Equivalent probabilistic control problem

At this point we are equipped to address the question from section III-D under the restriction of partial observability. To that end first acknowledge that problem (25) can be reinterpreted as the minimization of the moment projection between the posterior, $p_\rho(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T | \underline{z}_T)$, and the joint model, $p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T, \underline{z}_T)$. Further note that the resulting problem is equivalent to the moment projection of the posterior, $p_\rho(\underline{x}_T, \underline{u}_{T-1}, , \underline{y}_T, \underline{z}_T)$, and, the closed-loop density $p_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T)$. It follows that

$$\begin{aligned}(25) \equiv \arg & \min_{\underline{\pi}_{T-1}\in\mathcal{P}} \\ & \mathbb{D}\left[q_\rho(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T | \underline{z}_T) \parallel q_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T, \underline{z}_T)\right] \\ \equiv \arg & \min_{\underline{\pi}_{T-1}\in\mathcal{P}} \\ & \mathbb{D}\left[q_\rho(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T | \underline{z}_T) \parallel q_\pi(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T)\right]\end{aligned} \tag{27}$$

Again we may interpret the second problem as a probabilistic control problem with desired closed-loop density

$$q^*(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T) = q_\rho(\underline{x}_T, \underline{u}_{T-1}, \underline{y}_T | \underline{z}_T = \underline{\mathtt{true}}_T) \tag{28}$$

This observation now generalizes the equivalence between various probabilistic treatments of optimal control theory under the restriction of partial observability.

D. Dynamic programming and message passing on graphs

We finish the probabilistic treatment of POMDPs with a discussion on the use of the principle of dynamic programming, inspired by [10]. First note that (26) is equivalent to

$$\pi_t(u_t|w_t) = \frac{q_\rho(\underline{u}_t, \underline{y}_t, \underline{z}_T)}{q_\rho(\underline{u}_{t-1}, \underline{y}_t, \underline{z}_T)} \tag{29}$$

To reveal the corresponding message passing procedure we can focus on the probability $q_\rho(\underline{u}_t, \underline{y}_t, \underline{z}_T)$. This density can be decomposed as

$$\begin{aligned}q_\rho(\underline{u}_t, \underline{y}_t, \underline{z}_T) &= \int q_\rho(x_t, \underline{u}_t, \underline{y}_t, \underline{z}_T)\mathrm{d}x_t \\ &= \int q_\rho(x_t, w_t, \underline{z}_{t-1}) q_\rho(u_t, \overline{z}_t | x_t, w_t)\mathrm{d}x_t\end{aligned} \tag{30}$$

Next we focus on the density $q_\rho(u_t, \overline{z}_t | x_t, w_t)$. We have that

$$q_\rho(u_t, \overline{z}_t | x_t, w_t) = \rho_t(u_t|w_t) q_\rho(\overline{z}_t | x_t, w_t, u_t) \tag{31}$$

Then one can verify that

$$\pi_t(u_t|w_t) = \rho_t(u_t|w_t) \frac{\int \sigma_t(x_t) Q_t^\star(x_t, w_t, u_t)\mathrm{d}x_t}{\int \sigma_t(x_t) V_t^\star(x_t, w_t)\mathrm{d}x_t} \tag{32}$$

where

$$\begin{aligned}\sigma_t(x_t) &= q_\rho(x_t, w_t, \underline{z}_{t-1}) \\ Q_t^\star(x_t, w_t, u_t) &= q_\rho(\overline{z}_t | x_t, w_t, u_t) \\ V_t^\star(x_t, w_t) &= q_\rho(\overline{z}_t | x_t, w_t)\end{aligned} \tag{33}$$

It follows that (26) can be evaluated by means of a forward-backward message passing procedure. One easily verifies that the forward message, σ_t, satisfies the recursion

$$\begin{aligned}\sigma_t(x_t) = \rho_t(u_{t-1}|w_{t-1}) q(y_t) \int & p(x_t|x_{t-1}, u_{t-1}) \\ & \times q(z_{t-1}|x_{t-1}, u_{t-1}, y_{t-1}) \sigma_{t-1}(x_{t-1})\mathrm{d}x_{t-1}\end{aligned} \tag{34}$$

Further, the backward messages satisfy a similar recursions as in (16) and (17)

$$\begin{aligned}Q_t^\star(x_t, w_t, u_t) = q(z_t|x_t, u_t, y_t) \int & p(x_{t+1}|x_t, u_t) \\ & \times q(y_{t+1}) V_{t+1}^\star(x_{t+1}, w_{t+1})\mathrm{d}x_{t+1}\mathrm{d}y_{t+1}\end{aligned} \tag{35}$$

and

$$V_t^\star(x_t, w_t) = \int \rho_t(u_t|w_t) Q_t^\star(x_t, w_t)\mathrm{d}u_t \tag{36}$$

The result in (32) clearly generalizes the solution in (14) where the forward message becomes irrelevant on account of the full observability.

These results also rectify the result in [8] where a non-veridical joint density decomposition was entertained. By construction, there the posterior, $q_\rho(x_t|w_t)$, constituted a sufficient statistic. However, as is well-known in the context of POMPDs, this posterior is not a sufficient statistic when a veridical joint density decomposition is used [10].

VI. Discussion

In conclusion we give here some final remarks that we deem useful to gain further insight in the main results documented in this article. Further we aim to comment on the potential of the theory to solve practical control applications.

First we make a technical note considering the reward structure of c_t. Our results require that $p(z_t|x_t, u_t) \leq 1$. Since also $p(z_t|x_t, u_t) = \exp(c_t(x_t, u_t))$ it is implied that $c_t \leq 0, \forall t$. It is said that c_t has a positive reward structure. This is not a restrictive condition provided that c_t is bounded from above which poses a very reasonable assumption in practical applications. Further, provided that the exponential is convex the agent will be risk seeking. We appeal to section III-C for the utility theory interpretation of risk averse and risk seeking behaviour. It is emphasized that the present theory does not lend itself to encode risk averse behaviour.

The comment above can be brought into connection with the discussion on information-theoretic projections used in the probabilistic control framework, recall section III-D. The information- and moment projection are known to be *mode seeking* and *mode covering* [32, 33]. E.g. when the projections are used to approximate a heterogeneous Gaussian using a homogeneous Gaussian, the information-projected result will focus on the smallest eigenvalue of the heterogeneous covariance matrix whereas the moment-projected result will focus on the largest eigenvalue. As was shown in sections IV-C and V-C, the agents studied in this work are related to the moment projection. Put differently, the risk seeking agent attempts to cover the modes of the desired density. This behaviour can be avoided by considering the information-projection. Then the agent will be risk neutral and the probabilistic control framework will generate the maximum entropy RL objective. Unfortunately, then the agent can also no longer be determined by applying inference on a PGM.

We further argue that the main advantage of probabilistic control formulations is to be found in its mathematical and computational tractability. CaI and probabilistic control problems have often be praised for their efficacy to incite explorative behavioural tendencies [3, 7]. However, as noted by Millidge [25], the explorative behavioural tendencies obtained through CaI on MDPs boils down to entropy maximization of the policy. Rather the results in section IV-B and V-B, demonstrate that the problems treated in these earlier works isolate a problem from an iterative sequence that would converge to the optimal policy eventually.

Goal-directed exploration, as pursued by the dual control framework [34, 35], where an agent is learning its environment (i.e. identifying a veridical prescriptive model) while simultaneously attempting to execute a task, is an imminent feature of the POMDP framework when the learning dynamics are made explicit in the transition density. Computational challenges related to the solution of POMDPs have prevented further development and a wider spread reception of these endeavours. The fact that the solution of POMDPs can be approximated, or achieved iteratively, by solving an inference problem on a PGM could lead to new results in this area.

This leads us to provide some insights in the potential utility of our results related to practical algorithmic development. Already some ideas have been put forth in [6]. There the emphasis was on the (deterministic) trajectory optimization problem that could be solved by practising numerical tools from Bayesian estimation such as e.g. the extended Kalman smoother [36]. On the other hand, the majorisation property and the implication it has on the resulting policy sequence has important consequences for existing algorithms such as described in [18]. Further we acknowledge that it will remain challenging to forge the presented results into methods with practical utility in a partially observable setting. Though we may add here that inference problems on probabilistic graph models are easily automated, e.g. through message passing on factor graphs [37, 38].

Finally, the idea of maintaining a sequence of probabilistic policies that ultimately converges to an exact solution of an (PO)MDP also has an interesting biological interpretation. An agent maintains a PGM to keep track of its environment, then when a new task presents itself, the agent makes decisions by applying inference. The prior policy sequence, $\underline{\rho}_{T-1}$, encodes previous experience or any previously held beliefs to solve the task. The agent may then encode its new experience into the updated policy sequence, $\underline{\pi}_{T-1}$. In light of these final comment, we like to reiterate the surprising connection with AIF. One is lead to suspect that there are interesting contributions to be made by the control community to modern brain theories, possibly merging existing frameworks into a better unified understanding of human decision making.

Acknowledgements

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

Appendix A
The Expectation-Maximization algorithm

Consider a probabilistic model with hidden and observed variables, x and z. Further suppose that the probabilistic model, $\mathcal{M}$, is characterised by a set of variables θ, so that the joint density is given by, $p_\theta(z, x)$. One can then determine a Maximum Likelihood Estimation (MLE) of the parameters, θ, by maximizing the likelihood of the observations, z

$$\max_\theta \log \int p_\theta(z|x) p_\theta(x) \mathrm{d}x \tag{37}$$

It is well-known that it is difficult to treat this objective in a general setting. To circumvent the intractable inference, an auxiliary inference density, $q(x)$, can be introduced. The inference distribution allows to decompose the objective into a surrogate objective, $\mathcal{L}$, the so called evidence lower bound (ELBO), and, a relative entropy *error* term.

$$\begin{aligned}
&\log p_\theta(z) \\
&\quad = \int q(x) \log \frac{p_\theta(z, x)}{q(x)} \mathrm{d}x + \int q(x) \log \frac{q(x)}{p_\theta(x|z)} \mathrm{d}x \quad (38) \\
&\quad = \mathcal{L}[q(x)|z, \theta] + \mathbb{D}\left[q(x)||p_\theta(x|z)\right] \geq \mathcal{L}[q(x)|z, \theta]
\end{aligned}$$

Since the KL-divergence is positive semi-definite, with equality only if $q(x) \equiv p(x|z; \theta)$, the ELBO *minorizes* the log likelihood. Because of this property, the ELBO can be used to establish a Minorisation Maximization (MM) procedure. The MM principle denotes a general strategy to convert hard optimization problems into sequences of simple ones. The MM principle relies on a surrogate objective that is conditioned on the old parameters and that minorizes the true objective.

The surrogate is then used as a proxy of the true objective and minimized to find new parameters. The new parameters can be used to construct a new surrogate, and so forth. As such a sequence of optimization problems is established. In the context of the MLE problem this is referred to as the Expectation-Maximization or EM algorithm.

$$\theta^* \leftarrow \arg\max_{\theta} \mathbb{E}_{p(x|z;\theta^*)}[\log p(x, z; \theta)] \tag{39}$$

References

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

[2] M. Kárnỳ and T. V. Guy, "Fully probabilistic control design," *Systems & Control Letters*, vol. 55, no. 4, pp. 259–265, 2006.

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

[4] M. Toussaint and A. Storkey, "Probabilistic inference for solving discrete and continuous state markov decision processes," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 945–952.

[5] K. Rawlik, M. Toussaint, and S. Vijayakumar, "On stochastic optimal control and reinforcement learning by approximate inference," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[6] T. Lefebvre, "A review of probabilistic control and majorization of optimal control," 2022.

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

[8] T. Lefebvre, "Probabilistic majorisation of partially observable markov decision processes," in *Active Inference: Foruth International Workshop, IWAI 2023, Ghent, Belgium, September 13-15, 2023, Proceedings 4*. Springer, 2023.

[9] P. Whittle, *Optimal control : basics & beyond*. Wiley, 1996.

[10] M. R. James, J. S. Baras, and R. J. Elliott, "Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems," *IEEE transactions on automatic control*, vol. 39, no. 4, pp. 780–792, 1994.

[11] H. Kappen, "Linear theory for control of nonlinear stochastic systems," *Physical review letters*, vol. 95, no. 20, p. 200201, 2005.

[12] E. Todorov, "Linearly-solvable markov decision problems," in *Advances in neural information processing systems*, 2007, pp. 1369–1376.

[13] P. Dayan and G. E. Hinton, "Using expectation-maximization for reinforcement learning," *Neural Computation*, vol. 9, no. 2, pp. 271–278, 1997.

[14] H. Attias, "Planning by probabilistic inference," in *International workshop on artificial intelligence and statistics*. PMLR, 2003, pp. 9–16.

[15] D. Verma and R. P. Rao, "Goal-based imitation as probabilistic inference over graphical models," *Advances in neural information processing systems*, vol. 18, 2005.

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

[17] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint:1812.05905*, 2018.

[18] A. Abdolmaleki, J. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, "Maximum a posteriori policy optimisation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=S1ANxQW0b

[19] G. Williams, P. Drews, B. Goldfain, J. Rehg, and E. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1603–1622, 2018.

[20] T. Lefebvre and G. Crevecoeur, "Path integral policy improvement with differential dynamic programming," in *2019 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2019.

[21] ——, "Entropy regularised deterministic optimal control: From path integral solution to sample-based trajectory optimisation," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2022, pp. 401–408.

[22] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[23] 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.

[24] R. Smith, K. J. Friston, and C. J. Whyte, "A step-by-step tutorial on active inference and its application to empirical data," *Journal of mathematical psychology*, vol. 107, p. 102632, 2022.

[25] B. Millidge, A. Tschantz, A. K. Seth, and C. L. Buckley, "On the relationship between active inference and control as inference," in *Active Inference: First International Workshop, IWAI 2020, Co-located with ECML/PKDD 2020, Ghent, Belgium, September 14, 2020, Proceedings 1*. Springer, 2020, pp. 3–11.

[26] B. Millidge, A. Tschantz, and C. L. Buckley, "Whence the expected free energy?" *Neural Computation*, vol. 33, no. 2, pp. 447–482, 2021.

[27] L. Da Costa, N. Sajid, T. Parr, K. Friston, and R. Smith, "Reward Maximization Through Discrete Active Inference," *Neural Computation*, vol. 35, no. 5, pp. 807–852, 04 2023.

[28] J. Von Neumann and O. Morgenstern, "Theory of games and economic behavior, 2nd rev," 1947.

[29] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic control*, vol. 18, no. 2, pp. 124–131, 1973.

[30] P. Whittle, "Risk-sensitive linear/quadratic/gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.

[31] ——, "Risk sensitivity, a strangely pervasive concept," *Macroeconomic Dynamics*, vol. 6, no. 1, pp. 5–18, 2002.

[32] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[33] K. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] A. A. Feldbaum, "Dual control theory. i," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

[35] A. Feldbaum, "Dual control theory. ii," *Avtomatika i Telemekhanika*, vol. 21, no. 11, pp. 1453–1464, 1960.

[36] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, no. 3.

[37] 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.

[38] D. Bagaev, A. Podusenko, and B. de Vries, "Rxinfer: A julia package for reactive real-time bayesian inference," *Journal of Open Source Software*, vol. 8, no. 84, p. 5161, 2023.