Efficient Variational Sequential Information Control

Jianwei Shen
University of Arizona
Department of Computer Science

Jason Pacheco
University of Arizona
Department of Computer Science

Abstract

We develop a family of fast variational methods for sequential control in dynamic settings where an agent is incentivized to maximize information gain. We consider the case of optimal control in continuous nonlinear dynamical systems that prohibit exact evaluation of the mutual information (MI) reward. Our approach couples efficient message-passing inference with variational bounds on the MI objective under Gaussian projections. We also develop a Gaussian mixture approximation that enables exact MI evaluation under constraints on the component covariances. We validate our methodology in nonlinear systems with superior and faster control compared to standard particle-based methods. We show our approach improves the accuracy and efficiency of one-shot robotic learning with intrinsic MI rewards. Furthermore, we demonstrate that our method is applicable to a wider range of contexts, e.g., the active information acquisition problem.

1 INTRODUCTION

Optimal design is a fundamental problem in statistics that aims to choose a sequence of decisions that maximize some form of information gain or uncertainty reduction (Blackwell, 1950; Bernardo, 1979). Pioneering work by (Lindley, 1956) suggests a Bayesian approach that maximizes mutual information (MI) (Cover and Thomas, 2006; MacKay et al., 2003), in the context of Bayesian Optimal Experimental Design (BOED). This setting can be interpreted as minimizing expected posterior uncertainty over a fixed quantity of interest.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Yet, despite the apparent benefit of BOED, its practical application is limited by the difficulty of calculating the underlying MI utility (Paninski, 2003), which lacks a closed form in many nontrivial models. Numerous works exist for bounding and approximating MI using variational approximations (Poole et al., 2019), Monte Carlo approaches (Drovandi et al., 2013, 2014; Solonen et al., 2012), grid-based discretizations (Kim et al., 2014), and even explore these approximations in a BOED setting (Huan and Marzouk, 2016; Pacheco and Fisher, 2019; Foster et al., 2018, 2019; Kleinegesse and Gutmann, 2019).

This work extends variational BOED (Pacheco and Fisher, 2019; Foster et al., 2018, 2019) by modeling an evolving latent state driven by control inputs. Mutny et al. (2023) consider a similar discrete setting whereas our method applies to continuous latent states. In this way, our setting is more closely aligned with that of stochastic optimal control (Kushner et al., 2001; Bertsekas, 2012). We thus refer to our particular setting as *Optimal Information Control*. This setting has a wide range of applications, e.g. active simultaneous localization and mapping (Durrant-Whyte and Bailey, 2006; Stachniss et al., 2005; Carlone et al., 2014), active information acquisition (Atanasov et al., 2014; Charrow et al., 2014) and early work on childhood detection of social contingency (Movellan, 2005). Mutual information could also work as an intrinsic reward for RL tasks (Mohamed and Jimenez Rezende, 2015) when it lacks external rewards or rewards are sparse (Fischer and Tas, 2020).

Contributions. We address the computational aspects of control in this work by developing variational techniques for approximate inference and decision-making. We begin with an algorithm statement in the general case. Moving to Gaussian observation models we can show that our algorithm maintains a local upper bound on the MI reward. Finally, in models with Gaussian mixture model (GMM) dynamics, we present a constrained GMM projection that has a closed-form MI approximation. We numerically evaluate our methods in different experiments. In all cases, we observe

multiple orders of magnitude speedup over sequential Monte Carlo (SMC) with comparably superior accuracy.

2 PRELIMINARIES: MI CONTROL

We formulate optimal information control as an instance of stochastic control with mutual information (MI) rewards. Due to the intractable nature of the control problem, we provide a greedy approximation that maximizes instantaneous MI. However, unlike in standard control problems, the MI utility cannot be explicitly evaluated.

2.1 Optimal information control

We consider an optimal control problem for the dynamical system with latent variables $X_0^T = \{X_0, \dots, X_T\}$, observations $Y_1^T = \{Y_1, \dots, Y_T\}$, and joint PDF:

$$p(X_0^T, Y_1^T \,|\, d_1^T) = p(X_0) \prod_{t=1}^{T} p(X_t \,|\, X_{t-1}, d_t) p(Y_t \,|\, X_t) \quad (1)$$

where we denote the sequence of random variables and control inputs as X_0^T, Y_1^T, and $d_1^T = \{d_1, \dots, d_T\}$ respectively, and the rest of the paper shares the notation. Control inputs $d_t \in \mathcal{D}$ modulate the transition dynamics $p(X_t \mid X_{t-1}, d_t)$. The optimal information control problem is an instance of stochastic optimal control (Bertsekas, 2012), where we learn a policy π that optimizes cumulative mutual information over the sequence $t = 1, \dots, T$:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\ I(X_1^T; Y_1^T \mid \pi). \quad (2)$$

MI measures the dependence between random variables (Cover and Thomas, 2006), or equivalently expected reduction in entropy,

$$I(X; Y) = H(X) - \mathbb{E}_{p(y)}[H(X \mid Y = y)], \quad (3)$$

where entropy is given by the formula:

$$H(X) = \mathbb{E}[-\log p(X)], \quad (4)$$

and conditional entropy by:

$$H(X|y) = \mathbb{E}_{p(x|y)}[-\log p(X \mid y)]. \quad (5)$$

Solving this optimization problem in either open-loop (Atkinson et al., 2007; Ryan et al., 2016; Beck et al., 2018) or closed-loop (Huan and Marzouk, 2016; Drovandi et al., 2013; Solonen et al., 2012) manner is NP-hard in general (Bertsekas, 2012), necessitating a greedy approximation.

2.2 Sequential greedy information control

The global MI control objective in Eqn. (2) decomposes as a sum of conditional MI terms:

$$\max\ I(X_1; Y_1) + \sum_{t=2}^{T} I(X_t; Y_t \mid Y_1^{t-1}) \quad (6)$$

where we have dropped explicit dependence on the control policy to reduce notation. A derivation of the above is shown in the Appendix. A key property of Eqn. (6) is that each term depends on only a single latent state X_t. This suggests a simple greedy approximation at each time t:

$$\begin{aligned} d_t^* &= \underset{d}{\operatorname{argmax}}\ I(X_t; Y_t \mid \mathcal{H}_{t-1}, d) \\ &= \underset{d}{\operatorname{argmax}}\, H(X_t|\mathcal{H}_{t-1}, d) - H(X_t|Y_t, \mathcal{H}_{t-1}, d). \end{aligned} \quad (7)$$

Note that the conditioning set of the greedy objective Eqn. (7) is over observed measurements y_1^{t-1} in $\mathcal{H}_{t-1} = \{y_1^{t-1}, d_1^{t-1}\}$, as opposed to random variables Y_1^{t-1} as in the nonmyopic objective of Eqn. (6). This dependence on realized observations induces a closed-loop greedy sequential decision-making process.

The greedy MI reward (Eqn. (7)) is not directly observed and cannot be computed in most settings (Mafi et al., 2011; Still and Precup, 2012; Mazzaglia et al., 2022). A widely used MI approximation via a nested Monte Carlo (NMC) estimator is consistent, asymptotically unbiased, and admits a central limit theorem. On the other hand, it requires posterior samples and exhibits significant finite sample bias that decays slowly (Zheng et al., 2018; Rainforth et al., 2018) making them impractical in many settings. In Sec. 3, we propose our main contribution, an efficient variational approach that avoids costly sampling estimators.

3 VARIATIONAL MI CONTROL

This section provides details of our variational approach to information control in a general context. We start by motivating our approach with the difficulties of time-varying latent variables in the dynamic control problem. After that, we provide details of our approach including the use of *assumed density filtering (ADF)* and *expectation propagation (EP)* inference, and how these mechanisms yield variational MI approximations for control.

3.1 Variational MI Estimation

The control model of Sec. 2.1 incorporates decision controls that modulate dynamics via $p(X_t \mid X_{t-1}, d_t)$. Both entropy terms in the instantaneous (greedy) MI objective of Eqn. (7) involve the control variate d_t, and neither can be computed in closed-form. We re-

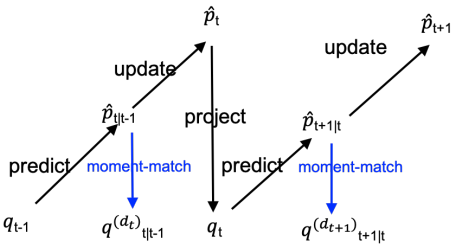


Figure 1: **Assumed density filtering and MI evaluation.** The traditional ADF updates (black lines) are adapted from Murphy (2012). Additional steps (blue) indicate projections for estimating MI in for decision d Eqn. (8).

place both terms with cross-entropies over variational distributions,

$$I(X_t; Y_t|\mathcal{H}_{t-1}, d_t) \approx H_{p_{d_t}}(q(X_t)) - H_{p_{d_t}}(q(X_t|Y_t)) \quad (8)$$

where

$$p_{d_t}(X_t \mid \mathcal{H}_{t-1}, d_t) \approx q(X_t \mid d_t) \quad (9)$$

and

$$p_{d_t}(X_t \mid Y_t, \mathcal{H}_{t-1}, d_t) \approx q(X_t \mid Y_t, d_t). \quad (10)$$

The approximation in Eqn. (8) was previously explored in the context of implicit likelihood models (Foster et al., 2019). Despite the simple form of the approximation in Eqn. (8), optimizing and evaluating it efficiently remains challenging, which we address next.

3.2 Greedy Variational MI control

The entropy terms in the variational approximation of instantaneous MI (Eqn. (8)) require expectations w.r.t. the joint posterior over state and measurement $p(X_t, Y_t \mid \mathcal{H}_{t-1}, d_t)$. This distribution is not available in practice so we perform approximate variational inference via *assumed density filtering (ADF)*, see Murphy (2012) for definition. Fig. 1 provides a depiction of the stages of ADF inference and MI approximation in our method.

Prediction Step Given a history of observations and decisions $\mathcal{H}_{t-1}$, we maintain an approximation of the posterior, $q_{t-1}(X_{t-1}) \approx \hat{p}_{t-1}(X_{t-1} \mid \mathcal{H}_{t-1})$, where $q_{t-1}(X_{t-1})$ is a member of the *exponential family*, e.g. a Gaussian distribution. For each hypothesized control variable d_t the prediction step computes an augmented distribution as,

$$\hat{p}_{t|t-1}(X_t, Y_t \mid \mathcal{H}_{t-1}, d_t) = p(Y_t \mid X_t) \int q_{t-1}(x_{t-1}) p(X_t \mid x_{t-1}, d_t)\, \mathrm{d}x_{t-1}. \quad (11)$$

This augmented distribution represents a local approximation to the predictive distribution and is not an exponential family in general so its entropy cannot be calculated easily.

Moment-matching Step Given the augmented distribution in Eqn. (11), the greedy MI surrogate objective $I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid \mathcal{H}_{t-1}, d_t)$ becomes,

$$\begin{aligned} I_{\hat{p}_{t|t-1}} &\approx H_{\hat{p}_{t|t-1}}(q_m(X_t)) - H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) \\ &\equiv I_{\hat{p}_{t|t-1}}(q), \end{aligned} \quad (12)$$

Cross entropy is an expectation w.r.t. the augmented distribution $\hat{p}_t$ instead of the true filter as in Eqn. (7). We approximate $\hat{p}_{t|t-1}(X_t \mid \mathcal{H}_{t-1}, d_t) \approx q_m$ and $\hat{p}_{t|t-1}(X_t \mid Y_t, \mathcal{H}_{t-1}, d_t) \approx q_c$. Finding the optimal variational distribution for Eqn. (12) is a non-convex optimization. We take the approach of Foster et al. (2019) and minimize an upper bound on the absolute error,

$$|I_{\hat{p}_{t|t-1}} - I_{\hat{p}_{t|t-1}}(q)| \le \min_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) + \min_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) + C \quad (13)$$

where C is a constant that does not affect the result of optimization. Standard approaches solved this by (stochastic) gradient descent. However, for Gaussian approximations $q^{(d_t)}_{t|t-1}(X_t, Y_t) = \mathcal{N}(m, \Sigma)$ we show that this bound can be efficiently solved via moment-matching. That this moment-matching step is optimal is not obvious, and is stated in the following theorem. For brevity, we drop explicit time indexing in the following theorem statement.

Theorem 3.1. *Let the joint $q(X, Y) = \mathcal{N}(m, \Sigma)$ match the moments of any target distribution $\hat{p}(X, Y)$. Then the marginal $q_m(X) = \int q(X, y)dy$ and conditional $q_c(X \mid Y) = q(X, Y)/q(Y)$ minimize the upper bound Eqn. (13).*

Dahlke et al. (2023) recently demonstrated a similar result for exponential families satisfying specific conditions. We provide a novel proof for the Gaussian case in the Appendix. We also show that the variational MI in Eqn. (12) takes a closed-form at the moment-matching solution.

Theorem 3.2. *Let the joint $q(X, Y) = \mathcal{N}(m, \Sigma)$ match the moments of any target distribution $\hat{p}(X, Y)$, then we have that $H_{\hat{p}}(q) = H_q(q)$.*

Update Step The projection step produces a set of variational approximations $q_{t|t-1}$ for each control parameter $d_t \in \mathcal{D}$. MI is empirically evaluated via Eqn. (12) for each of these quantities and the control d_t with maximum (approximate) MI is selected. The chosen control is executed and a realized measurement is obtained $y_t \sim p(Y_t \mid X_t)$ from the environment. Finally, an ADF update is performed to yield an exponential family approximation $q_t \approx \hat{p}_t(X_t \mid \mathcal{H}_t)$ via KL-projection (e.g. moment-matching in the exponential family).

Expectation propagation (EP) inference.

The ADF-driven variational approach outlined in Sec. 3.2 and Fig. 1 has the benefit of being computationally efficient as inference proceeds in a single forward-pass. However, the non-iterative nature of ADF can lead to poor results in some cases (Minka, 2001). To address this, we consider EP as an alternative inference approach in the **update step**.

Our implementation follows standard EP procedures (Minka, 2001; Heskes and Zoeter, 2002; Seeger, 2005). See the Appendix for algorithm pseudocode. EP tends to yield more accurate inference than ADF but does pose some practical drawbacks. First, there is additional computational overhead due to iterative message updates. Second, it can exhibit non-convergence or numerical instability (Minka, 2004; Wainwright et al., 2008). Using standard techniques to avoid these drawbacks (message damping and abandoning numerically invalid updates) we find that our EP implementation is reliable in practice (see experiments Sec. 6).

4 MI CONTROL FOR GAUSSIAN NOISE SYSTEMS

We consider a class of systems with linear Gaussian observation likelihood: $p(Y_t \mid X_t) = \mathcal{N}(Y_t \mid AX_t + a, B)$, and apply our general approach to this model with Gaussian approximation. After that, we introduce a special Gaussian noise system, the GMM-Gaussian system, which has generalizability to other complex models. Within this system, we propose the development of a constrained Gaussian mixture variational family in that allows analytic calculation of MI under the variational mixture projection.

4.1 Gaussian MI approximation

We apply ADF and EP as discussed in Sec. 3.2 with Gaussian variational approximations. Following the **prediction step** in Sec. 3.2 we have an upper bound of the MI w.r.t the augmented distribution.

Theorem 4.1. *In a model with linear-Gaussian observations, given the augmented distribution at time t,* $\hat{p}_{t|t-1}(X_t, Y_t \mid \mathcal{H}_{t-1}, d_t)$ *and its Gaussian approximation,* $q^{(d_t)}_{t|t-1}(X_t, Y_t)$*, by moment-matching,*

$$I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid \mathcal{H}_{t-1}, d_t) \leq I_{q^{(d_t)}_{t|t-1}}(X_t; Y_t \mid \mathcal{H}_{t-1}, d_t)$$

See proof in the Appendix. This theorem holds independent of the state dynamical transitions, provided the measurement model is a linear Gaussian distribution. Note that the bound is w.r.t. the local approximation under the augmented distribution as opposed to the true filter. To better adapt to the true GMM filter distribution we develop a specialized approach based on GMM projections outlined next.

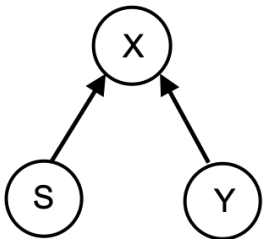


Figure 2: **PGM** of constrained GMM

4.2 Constrained GMM approximation

We consider a general class of GMM-Gaussian control model, i.e.,

$$p(X_0^T, Y_1^T \mid d_1^T) = \mathcal{N}(X_0 \mid m_0, \Sigma_0) \prod_{t=1}^{T} \mathcal{N}(Y_t \mid FX_t, R) \sum_{k=1}^{K} w_{k,d_t} \mathcal{N}(X_t \mid A_{k,d_t} X_{t-1}, Q_{k,d_t}). \quad (14)$$

The prior on X_0 and observation likelihood are both given by standard linear Gaussian distributions. Nonlinearities in the model are introduced by the GMM transition dynamics. Under control $d_t \in \mathcal{D}$ transitions are K-component GMMs with weights $\sum_{k=1}^{K} w_{k,d_t} = 1$ and components $\mathcal{N}(X_t \mid A_{k,d_t} X_{t-1}, Q_{k,d_t})$. The filter distribution at time t is a GMM with $\mathcal{O}(K^t)$ components, making inference NP-hard. Besides, this model also generalizes since the Gaussian mixture is considered a universal density approximator (Maz'ya and Schmidt, 1996) and linear Gaussian observations can easily be extended to a Gaussian mixture model, making this a good candidate for general study. Moreover, such models are widely used for model-based learning of dynamical systems (Khansari-Zadeh and Billard, 2011; Hersch et al., 2008). At each time $t-1$, we maintain an ensemble of K Gaussians, $q_{t-1}(X_{t-1}, S_{t-1} = k) =$

$$\pi_{k,t-1} \mathcal{N}(X_{t-1} \mid \hat{\mu}_{k,t-1}, \hat{\Sigma}_{k,t-1}), \quad (15)$$

where S_{t-1} is a discrete random variable, $q(S_{t-1} = k) = \pi_{k,t-1}$. Note that a fixed-component Gaussian ensemble is in the exponential family, allowing ADF/EP updates via moment-matching (Heskes and Zoeter, 2002; Pacheco and Sudderth, 2012). Following is the breakdown of the steps.

Prediction Step The GMM variational approximation in Eqn. (15) yields a K^2-component GMM augmented distribution:

$$\hat{p}_{t|t-1}(X_t, Y_t, S_t = k \mid \mathcal{H}_{t-1}, d_t) = w_k \mathcal{N}(Y_t \mid FX_t, R) \, q_{t-1}(X_{t-1}) \mathcal{N}(X_t \mid A_{k,d_t} X_{t-1}, Q_{k,d_t}) \quad (16)$$

Constrained GMM Moment-matching Step For each control variate we project the K^2 GMM augmented distribution of Eqn. (16) to a K-component

Gaussian ensemble via KL-projection,

$$q^{(d_t)}_{t|t-1}(X_t, Y_t, S_t = k) = \operatorname*{argmin}_q \mathrm{KL}(\hat{p}_{t|t-1}(X_t, Y_t, S_t = k \mid \mathcal{H}_{t-1}, d_t) \,\|\, q) \quad (17)$$

However, the entropy (or cross-entropy) of a GMM lacks a closed-form in general (Huber et al., 2008). We, therefore, restrict the form of our projection onto an ensemble with factorization (see Fig. 2 for PGM),

$$q^{(d_t)}_{t|t-1}(X_t, Y_t, S_t = k) = \omega_k \mathcal{N}(Y_t \mid \eta, P)\mathcal{N}(X_t \mid F_k Y_t, M_k). \quad (18)$$

In particular, the marginal moments of Y_t are invariant to the mixture component, whereas marginal and conditional moments of X_t are component-dependent. Therefore, η and P in Eqn. (18), defined as the mean and covariance of the proposal distribution, can be directly computed from the target distribution. The optimal F_k and M_k are found by solving,

$$\operatorname*{argmin}_{F_k, M_k} \mathbb{E}_{\hat{p}_{t|t-1}}[-\log q^{(d_t)}_{t|t-1}(X_t \mid Y_t, S_t = k)]. \quad (19)$$

Considering that $q^{(d_t)}_{t|t-1}(X_t \mid Y_t, S_t = j)$ is a Gaussian, the function is convex and can be solved analytically. The solution is efficient and more importantly leads to an analytic expression for the following MI: $I_{q^{(d_t)}_{t|t-1}}(\{X_t, S_t\}; Y_t)$. We show this calculation briefly via repeated application of the entropy chain rule (Cover and Thomas, 2006):

$$I_{q^{(d_t)}_{t|t-1}}(\{X_t, S_t\}; Y_t) = \sum_{k=1}^{K} \omega_k [H_{q^{(d_t)}_{t|t-1}}(X_t \mid S_t = k) - H_{q^{(d_t)}_{t|t-1}}(X_t \mid Y_t, S_t = k)] \quad (20)$$

The result is a mixture of marginal and conditional Gaussian entropies, each of which has a closed form. The resulting estimator appears similar to but differs from, the application of Jensen's inequality to compute GMM entropy. Fig. 3a shows that the constrained GMM MI estimator compares comparably to other proposed MI estimates, and substantially outperforms Jensen's. Note that by non-negativity of MI and an application of the chain rule, we have a useful upper bound,

$$I_{q^{(d_t)}_{t|t-1}}(\{X_t, S_t\}; Y_t) \geq I_{q^{(d_t)}_{t|t-1}}(X_t; Y_t) \quad (21)$$

We thus use MI of the constrained GMM projection $I_{q^{(d_t)}_{t|t-1}}$ (Eqn. (20)) as an approximation of $I_{\hat{p}_{t|t-1}}$. Our surrogate does not bound that quantity nor the desired true posterior MI $I_{p_{t|t-1}}$, but performs well empirically (c.f. Fig. 3a). We further demonstrate the constrained GMM projection yields a reasonable approximation for a GMM distribution to mimic the filter distribution, by conducting a standalone experiment. This experiment shows two facts in a 10-decision-making problem in a single-step scenario: both Theorem 4.1 and Eqn. 21 hold; our methods (red and yellow bars) not only estimate the MI of a bivariate GMM close to the "ground truth" (blue bars), estimated by the numerical approximation, but also produce high-quality decisions, tagged with diamond characters. Jensen's inequality provides a rough estimation of the MI, which we use as a baseline comparison. The MI estimations by different methods are illustrated in Fig. 3a.

5 RELATED WORK

We position our work in the context of three research areas:

BOED is a simplification of our present setting where the latent quantity X is static. Our work extends the variational BOED approaches of Pacheco and Fisher (2019); Dahlke et al. (2023); Foster et al. (2019) to the case where X_t evolves over time. Alternatives to variational BOED include sample-based methods that rely on a nested Monte Carlo (NMC) estimator (Zheng et al., 2018; Rainforth et al., 2018) and for implicit models (Kleinegesse and Gutmann, 2019). Variations of BOED exist with time-varying X_t, but where the decision variable modulates only the observation model (Williams, 2007; Shamaiah et al., 2010).

Stochastic dynamic control. Our work can be viewed as an instance of stochastic dynamic control (Bertsekas, 2012) with some modifications. Firstly, it is a partially observed Markov decision process (POMDP), secondly, the MI reward is not explicitly observed and lacks a closed form. Discretization methods have been proposed in this setting (Huan and Marzouk, 2016), but have only been demonstrated for a static (BOED-style) setting and do not scale. Our work most closely resembles the setting of Mutny et al. (2023), but whereas they restrict to discrete latent states we consider continuous states and observations.

Active Information Acquisition subsumes several problem areas in RL and control where there exists an implicit information reward. One such area is active sensing, where a remote sensor gathers information (Ryan and Hedrick, 2010; Atanasov et al., 2014) to map the environment (Bennetts et al., 2013), perform search and rescue (Kumar et al., 2004), or surveillance (Rybski et al., 2000). Similarly, intrinsically motivated RL includes an implicit motivation of information gathering, often in concert with an extrinsic reward Oudeyer and Kaplan (2008). A commonly used utility function, *empowerment*, measures MI between actions and states (Salge et al., 2014). Our work more closely aligns with that of Houthooft et al. (2016) in that we consider MI between states and observations. A variant that considers static latent states can be used

to formulate active SLAM (Durrant-Whyte and Bailey, 2006).

6 EXPERIMENTAL RESULTS

The plausibility and efficiency of the proposed methods are validated by three experiments. All of these experiments suggest that our methods are applicable in the online control setting without an extra offline training phase. We compare our methods with the particle-based method, i.e., Particle Filtering (PF) for the inference, and Nested Monte Carlo (NMC) for the MI estimation. The NMC is widely used as a comparison baseline for the MI estimation, e.g., Foster et al. (Foster et al., 2018). The first experiment is conducted in a GMM-Gaussian dynamical system, defined in Sec. 4. We compare our approaches with the particle-based method in terms of efficiency, accuracy, and decision quality. In the second experiment, we apply the MI reward to a variant of the robotic problem (Porta et al., 2006) in the Partially Observable Markov Decision Process (POMDP) and show that our methods are applicable in the scenario with sparse explicit reward. And it will improve the performance in efficiency and accuracy in the one-short learning style. Finally, we demonstrate our methods are widely applicable by applying them to a problem of active information acquisition, which lacks an explicit reward function. Further details are included from Sec. 6.1 to Sec. 6.3 correspondingly. We conduct these experiments on an iMac with a 3-GHz 6-Core Intel i5 processor and 32 GB memory.

6.1 GMM-Gaussian system

We validate our method on the GMM-Gaussian dynamical system of Eqn. (14) over a horizon of $T = 20$ time steps and a choice of 5 transition models. We compare the results of six different methods to make decisions for optimizing the MI: numerical approximation, random choice, Particle Filtering (PF), ADF-Gaussian (ADF), EP-Gaussian (EP), and ADF-GMM. To evaluate cumulative MI against optimal (greedy) selection and inference we use a numerical Riemann sum approximation on a grid size of $6,000$ bins in each dimension.

After the decision-making process, we evaluate the cumulative information of trajectories relative to the optimal numerical sequence. Given the decision trajectory $\mathcal{T}$ of a method and the measurements it received, we compute the cumulative MI using numerical Riemann sum integration as $\sum_t H(X_t \mid \mathcal{H}_{t-1}) - H(X_t \mid Y_t, \mathcal{H}_{t-1})$. Accuracy is scored relative to the trajectory generated by numerical inference and optimal greedy decision-making, which we use as ground truth.

Comparable accuracy to PF baseline. We conduct 11 independent trials of each method. For each trial, we evaluate the realized information gain relative to optimal as previously described. Fig. 3c reports the mean cumulative MI at each state and +/- 1 STDEV relative to optimal numerical selection. In all cases, the mean results are better than the baseline PF with $1,000$ particles. We note that variational methods demonstrate significantly tighter standard deviations of error. All methods outperform random selection.

Orders of magnitude speedup. To evaluate performance we record the running time for making a decision at each step t and calculate the average performance. We experiment on multiple combinations of n-component GMM dynamic transition and k decisions, where $n \in \{2, 3, 4, 5, 6, 7, 8\}$ and $k \in \{5, 6, 7, 8, 9, 10\}$. Fig. 3d demonstrates that our variational methods are orders of magnitude faster than PF as a function of the number of decisions at each time point, varying from 5 to 10. ADF methods show 4 orders of magnitude speedup while the iterative overhead of EP yields 2 orders of magnitude. However, we note that EP has tighter confidence intervals in Fig. 3c as a result. The ADF-GMM inference shows minimal overhead compared to Gaussian ADF. See similar speedups as a function of the number of components in the GMM transition in the appendix.

Qualitatively comparable inference to baseline. Our methods produce qualitatively similar inferences compared to PF, but we reiterate that the variational approaches are orders of magnitude faster. To save space, we put them into the appendix.

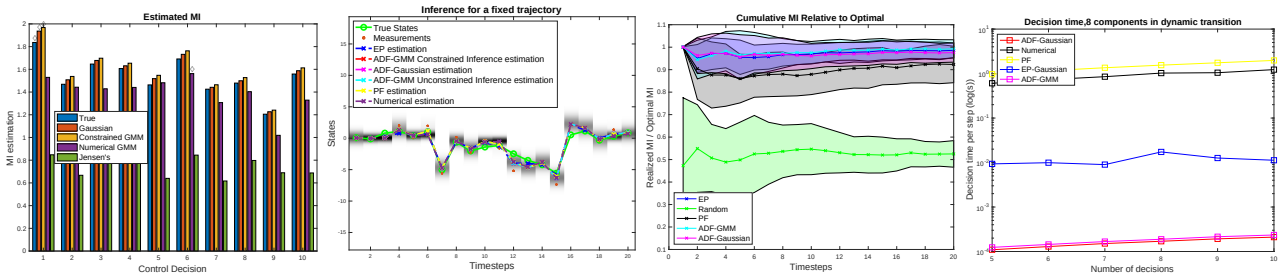
6.2 Continuous POMDP

Environment and Challenge. We adopt the basic environmental setup as Porta et al. (2006) but focus on a one-shot POMDP learning problem to underline the need for **efficiency** and **accuracy**. The robot's objective is to reach the correct door by navigating the corridor (taking actions to move left, right, and enter), as Fig. 4a. Operated in the POMDP setting, the robot's **continuous** true positions $X_t \in [-21, 21]$ are latent, but it receives **continuous** noisy measurements $Y_t \in [0, 5]$ of the width of the current position. The state dynamic transition is a linear Gaussian distribution dependent on three decisions,

$$p(X_{t+1} \mid X_t, d_t) = \mathcal{N}(X_t + a_d, 0.05), \tag{22}$$

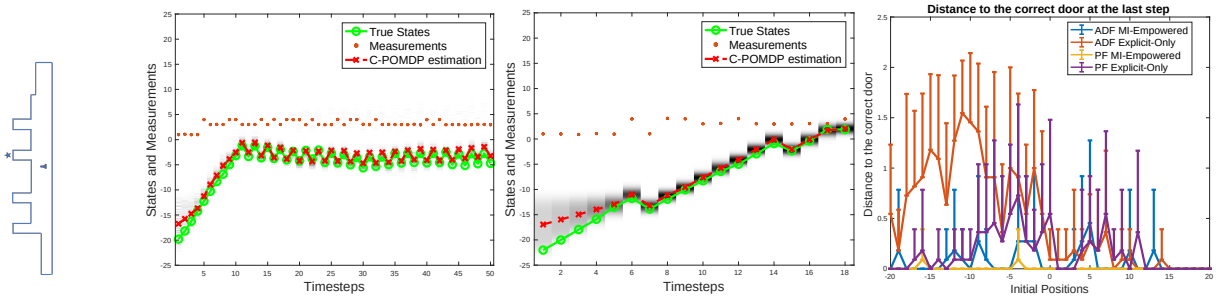
where $a_d \in \{0, -2, +2\}$. The measurement model is defined as

$$p(Y_t \mid X_t) \propto \sum_{i=1}^{n} \mathcal{N}^{-1}(X_t \mid \eta_i, \Lambda_i)\mathcal{N}(Y_t \mid \mu_i, \Sigma_i), \tag{23}$$



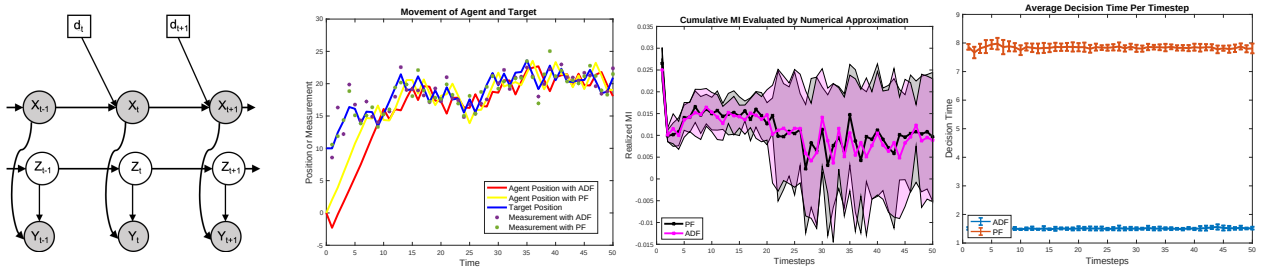
(a) MI estimations (b) Posterior inference (c) MI gained (d) Runtime

Figure 3: (a) This figure demonstrates the MI estimations of the Gaussian mixture model with $K^2 = 16$ components by five methods. The marked bars with a diamond character represent the decisions they will take respectively. All our methods always outperform Jensens' inequality estimation (green bars) in terms of making a decision. (b) Posterior inference of the same trajectory for different methods, grayscale bars representing the filter distribution approximated by the numerical approximation method. (c) This figure shows the cumulative MI by each method (with different decision trajectories chosen) versus the optimal MI. We demonstrate the ratio of these two values to validate that our approaches, ADF-Gaussian, EP, and ADF-GMM, will yield high-quality decisions. (d) We fix the number of components in the GMM dynamic transition to 8 and collect the average running time of each method for making a decision. The number of decisions is from 5 to 10. ADF-Gaussian and ADF-GMM are much faster than the rest, and EP-Gaussian also achieves a reliable performance in running time.



(a) Environment (b) Explicit only reward (c) MI-empowered reward (d) The last distance

Figure 4: (a) This is a demonstration of the environment. (b) It shows a case where the exact-reward-only mode can cause the oscillation of the robotic motion and lead to getting stuck. We approximate the true belief states by PF and the MI is approximated by the NMC estimator in this figure. (c) This figure shows a case where we apply the MI-empowered reward and ADF inference and decision-making process. From the same initial position as the (b), it explores the surroundings first and moves towards the target door after it is quite certain about its position. (d) It represents the distance between the robot and the mean value of the correct door at the last step. Our ADF method with an MI-empowered reward achieves almost the same accuracy as the near-optimal PF approximation with an MI-empowered reward. And it is more stable than the explicit-only methods.



(a) PGM. (b) Movement (c) MI (d) Runtime

Figure 5: (a) The PGM. The agent (X) aims to reach the stochastically moving target (Z) by dynamically transmitting its position according to the measurement (Y). (b) The trajectory taken by the agent, applying two different methods, with the true target position and measurement received. (c) The average realized MI is estimated by numerical approximation given the z_1^T, x_1^T and y_1^T for 11 repeated trials. Both methods achieve similar MI. (d) The average runtime at each time step for 11 repeated trials. Our method is approximately 5x faster than the PF.

where η_i and Λ_i are natural parameters of the Gaussian distribution. We plot the joint distribution in Fig. 6a. The environment is composed of four different features, left edge, right edge, corridor, and doors with different widths, {1,2,3,4}. We plot the distribution of different positions given the feature, i.e., $p(X \mid Y = y)$, where $y \in \{1, 2, 3, 4\}$, in Fig. 6b. The explicit reward function plotted in Fig. 6c implies that it is sparse in the given domain, which suggests a slow convergence and can use empowerment in online few-shots learning settings. This simulated environment can be extended to real-world scenarios like fire rescue, where the robot needs efficient localization and target-finding capabilities in unknown environments.

Computing the *belief state*, i.e. the posterior distribution $b_{t-1} = p(X_{t-1} \mid \mathcal{H}_{t-1})$, is infeasible since it is a GMM and the number of components grows exponentially. Given the belief state b_{t-1} and reward function $r(a, X)$, the optimal policy is learned by

$$\pi^* = \underset{\pi}{\operatorname{argmax}} E_{p(X_t, Y_t \mid \mathcal{H}_{t-1})}[r(\pi(x), X_t = x) + \bar{R}], \quad (24)$$

where $\bar{R}$ is the future reward. Learning an optimal policy is NP-hard, even its approximation is computationally prohibitive. A straightforward and effective method is to apply a greedy reward. But we observe that, with $E_{p(X_t, Y_t \mid \mathcal{H}_{t-1})}[r(a_t, X_t = x)]$ (greedy expected explicit reward) only, the performance of the robot is not stable and it tends to get stuck at one place as Fig. 4b.

MI-empowered reward. Inspired by intrinsically motivated RL and curiosity-driven RL, we modify the reward function by adding an MI term between the latent state and measurement,

$$R(a_t) = E_{p(X_t, Y_t \mid \mathcal{H}_{t-1})}[r(a_t, X_t = x)] + \alpha I(X_t; Y_t \mid a_t, \mathcal{H}_{t-1}). \quad (25)$$

The α value is a 0-1 value set to control the balance of exploration and exploitation, i.e. the reward encourages the robot to explore more when it is not certain about its position but prevents the robot from overly exploring when it has a near-precise belief of its position. In practice, we set it to 0 when the variance of the belief state is below a threshold $\gamma = 0.5$.

Methodology. We approximate the belief state by a fixed-number Gaussian ensemble by moment-matching and apply a Gaussian MI approximation shown in Sec. 4.1 to estimate the MI term in Eqn. (25). As a comparison, we also implement PF with 3000 samples in this space to approximate the truth. As shown in Fig. 4c, with MI-empowered reward, the robot has the ability to address the oscillation problem in Fig. 4b. In our experiment, the robot first explores the area for self-localization and then it moves to the target area when it is equipped with the MI-empowered reward.

Comparable accuracy and significant improvement in speeding up the process. We terminate the process once the robot enters within the range of the correct door. To assess the method's accuracy, we collect the distance to the correct door at the last step (Fig. 4d) in 11 runs and calculate the mean and $\pm$ 1 STDEV. The result confirms the accuracy of our method (ADF MI-empowered) is nearly as accurate as the PF with MI-empowered reward and outperforms explicit-reward-only methods. Detailed computational speed advantages are discussed in Sec. 6.1. See other experimental results in the appendix.

6.3 Active information acquisition

We apply our method in an active information acquisition context (Atanasov et al., 2014; Charrow et al., 2014). We illustrate the PGM of this problem in Fig. 5a, with **observed** agent states X_1^T controlled by d_1^T, stochastically moving **latent** target states Z_1^T, and **observable** measurements Y_1^T, related with both the agent and the target. The goal of the agent is to maximize information $I(Z_1^T; Y_1^T \mid X_1^T)$. Observation noise scales with the relative distance to target, so this objective results in the agent acquiring the target.

Model and MI objective. The agent moves by choosing from three discrete decisions to the left, to the right, and stays, with a Gaussian noise for all the cases,

$$p(X_{t+1} \mid X_t, a_d) = \mathcal{N}(X_{t+1} \mid X_t + a_d, \sigma^2), \quad (26)$$

where a_d is a control variate. The target moves stochastically by the model

$$p(Z_{t+1} \mid Z_t) = \sum_i^K w_i \mathcal{N}(Z_{t+1} \mid Z_t + b_i, \Sigma_i). \quad (27)$$

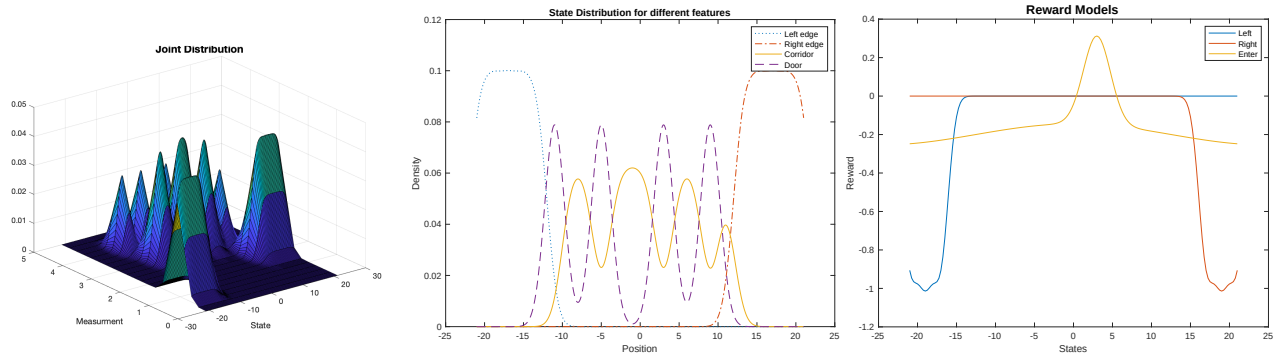
At each time, the agent receives a measurement based on the position of its own and the target,

$$p(Y_t \mid Z_t, X_t) = \mathcal{N}(Y_t \mid Z_t, \alpha |Z_t - X_t| + \epsilon). \quad (28)$$

Without an explicit reward objective, we apply the mutual information $I(Z_t; Y_t \mid X_t, d_t, \mathcal{H}_{t-1})$ as the intrinsic reward per step. To note, $\mathcal{H}_{t-1} = \{y_1^{t-1}, x_1^{t-1}, d_1^{t-1}\}$, includes realizations of agent states, decisions, and measurements of the past. This reward motivates the agent to move towards the target and keep track of it.

Difficulty and methodology. The biggest hurdle originates from the exact inference of the belief state. Compared to the previous experiments, the exact belief state is not only intractable but also has no closed-form solution. It also causes difficulty in MI calculation since both terms have no closed-form solution.

$$\max_{d_t} H(Y_t \mid X_t, d_t, \mathcal{H}_{t-1}) - H(Y_t \mid Z_t, X_t, d_t, \mathcal{H}_{t-1}), \quad (29)$$



(a) The joint distribution. (b) The state distribution given features. (c) The reward model of the experiment.

Figure 6: (a) This figure shows the joint distribution of states and measurements, which displays the multi-modality of this modal for most of the cases. (b) It demonstrates the distribution of states given mean values of measurements for four features, i.e., $p(X \mid Y = y)$, where $y \in \{1, 2, 3, 4\}$ and they represent *Left edge*, *Right edge*, *Corridor*, and *Door* respectively. (c) This plot depicts the explicit reward values for different decisions evaluated at different states, where the positive values only occur when the robot **enters** within the range of the correct door at position 3.

Thus, we record the approximate belief state

$$p(Z_{t-1} \mid \mathcal{H}_{t-1}) \approx \mathcal{N}(Z_{t-1} \mid \mu_{t-1}, \Sigma_{t-1}). \quad (30)$$

During the decision-making phase, we compute the augmented latent state marginal distribution $\hat{p}(Z_t \mid \mathcal{H}_{t-1})$ with the approximate belief state and draw samples $\{Z_t^{(s)}\}_{s=1}^N$ from it. Independently, we forward sample the agent state $\{X_t^{(s)}\}_{s=1}^N$ from $p(X_t \mid X_{t-1} = x_{t-1}, d_t)$. For each sample $X_t^{(s)}$, we approximate

$$\hat{p}(Y_t \mid X_t^{(s)}) \approx \frac{1}{N} \sum_s^N p(Y_t \mid Z_t^{(s)}, X_t^{(s)}), \quad (31)$$

and estimate its entropy by moment-matching as discussed before. To obtain the updated belief state after receiving a new measurement, we apply the ADF. As a comparison, we also implement the PF and estimate the MI by NMC.

Comparable accuracy and significant speedup. We evaluate our method from three dimensions, the completion of reaching and tracking the target, the realized MI estimated by the numerical approximation, and the time for making decisions. We compare our method with PF. We discover that our method has a similar performance in task completion and MI acquisition, but achieves a significant boost in runtime. See results from Fig. 5b to Fig. 5d.

7 LIMITATIONS

Methods proposed follow similar limitations to that of ADF and EP. In particular, moments of the augmented distribution must be computable either analytically or empirically. In general the augmented distribution in Eqn. (11) should have a closed-form or be approximable by numerical methods. Our experiments assume a known dynamics and observation model, though model-based learning is possible but beyond the present scope.

8 CONCLUSION

We propose an efficient approach that combines fast variational inference with variational MI approximation in a cohesive framework for information control. We demonstrate the effectiveness of our approach in the robotic control and the active information acquisition context. We believe our method is well-suited for these problems and plan to explore them in future work. Note that greedy decision-making can be arbitrarily suboptimal and closed-loop optimal methods provide quality guarantees, see Williams (2007) for reference, which is the focus of future work.

References

D. Blackwell. Comparison of experiments. In Jerzy Neyman, editor, *2nd BSMSP*, pages 93–102, Berkeley, CA, August 1950. UC Berkeley.

J. M. Bernardo. Expected Information as Expected Utility. *Ann. Stat.*, 7(3):686–690, May 1979.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

Christopher C Drovandi, James M McGree, and Anthony N Pettitt. Sequential monte carlo for bayesian sequentially designed experiments for discrete data. *Computational Statistics & Data Analysis*, 57(1):320–335, 2013.

Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A sequential monte carlo algorithm to incorporate model uncertainty in bayesian sequential design. *Journal of Computational and Graphical Statistics*, 23(1):3–24, 2014.

Antti Solonen, Heikki Haario, and Marko Laine. Simulation-based optimal design using a response variance criterion. *Journal of Computational and Graphical Statistics*, 21(1):234–252, 2012.

Woojae Kim, Mark A Pitt, Zhong-Lin Lu, Mark Steyvers, and Jay I Myung. A hierarchical adaptive approach to optimal experimental design. *Neural computation*, 26(11):2465–2492, 2014.

Xun Huan and Youssef M Marzouk. Sequential bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320*, 2016.

Jason Pacheco and John Fisher. Variational information planning for sequential decision making. pages 2028–2036, 2019.

Adam Foster, Martin Jankowiak, Eli Bingham, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational optimal experiment design: efficient automation of adaptive experiments. 2018.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32, 2019.

Steven Kleinegesse and Michael U Gutmann. Efficient bayesian experimental design for implicit models. pages 476–485, 2019.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*, pages 7349–7374. PMLR, 2023.

Harold Joseph Kushner Kushner, Harold J Kushner, Paul G Dupuis, and Paul Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer Science & Business Media, 2001.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.

Cyrill Stachniss, Giorgio Grisetti, and Wolfram Burgard. Information gain-based exploration using rao-blackwellized particle filters. In *Robotics: Science and systems*, volume 2, pages 65–72, 2005.

Luca Carlone, Jingjing Du, Miguel Kaouk Ng, Basilio Bona, and Marina Indri. Active slam and exploration with particle filters using kullback-leibler divergence. *Journal of Intelligent & Robotic Systems*, 75:291–311, 2014.

Nikolay Atanasov, Jerome Le Ny, Kostas Daniilidis, and George J Pappas. Information acquisition with sensing robots: Algorithms and error bounds. In *2014 IEEE International conference on robotics and automation (ICRA)*, pages 6447–6454. IEEE, 2014.

Benjamin Charrow, Vijay Kumar, and Nathan Michael. Approximate representations for multi-robot control policies that maximize mutual information. *Autonomous Robots*, 37:383–400, 2014.

Javier R Movellan. An infomax controller for real time detection of social contingency. In *Proceedings. The 4th International Conference on Development and Learning, 2005*, pages 19–24. IEEE, 2005.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

Johannes Fischer and Ömer Sahin Tas. Information particle filter tree: An online algorithm for pomdps with belief-based rewards on continuous domains. In *International Conference on Machine Learning*, pages 3177–3187. PMLR, 2020.

Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*, volume 34. OUP Oxford, 2007.

Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

Joakim Beck, Ben Mansour Dia, Luis FR Espath, Quan Long, and Raul Tempone. Fast bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, 334:523–553, 2018.

Nassim Mafi, Farnaz Abtahi, and Ian Fasel. Information theoretic reward shaping for curiosity driven learning in pomdps. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–7. IEEE, 2011.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Pietro Mazzaglia, Ozan Catal, Tim Verbelen, and Bart Dhoedt. Curiosity-driven exploration via latent bayesian surprise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7752–7760, 2022.

Sue Zheng, Jason Pacheco, and John Fisher. A robust approach to sequential information theoretic planning. In *International Conference on Machine Learning*, pages 5936–5944, 2018.

Tom Rainforth, Robert Cornish, Hongseok Yang, and Andrew Warrington. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4264–4273, 2018.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Caleb Dahlke, Sue Zheng, and Jason Pacheco. Fast variational estimation of mutual information for implicit and explicit likelihood models. In *International Conference on Artificial Intelligence and Statistics*, pages 10262–10278. PMLR, 2023.

Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369, 2001.

TM Heskes and OR Zoeter. Expectation propagation for approximate inference in dynamic bayesian networks. In *Darwiche, A.; Friedman, N.(eds.), Uncertainty in artificial intelligence: proceedings of the eighteenth conference (2002), August 1-4, 2002, University of Alberta, Edmonton*, pages 216–233. San Francisco, Calif.: Morgan Kaufmann Publishers, 2002.

Matthias Seeger. Expectation propagation for exponential families. Technical report, 2005.

Thomas Minka. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Vladimir Maz'ya and Gunther Schmidt. On approximate approximations using gaussian kernels. *IMA Journal of Numerical Analysis*, 16(1):13–29, 1996.

S Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27 (5):943–957, 2011.

Micha Hersch, Florent Guenter, Sylvain Calinon, and Aude Billard. Dynamical system modulation for robot learning via kinesthetic demonstrations. *IEEE Transactions on Robotics*, 24(6):1463–1467, 2008.

Jason L Pacheco and Erik B Sudderth. Improved variational inference for tracking in clutter. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 852–855. IEEE, 2012.

Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188. IEEE, 2008.

Jason L Williams. Information theoretic sensor management. 2007.

Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *49th IEEE conference on decision and control (CDC)*, pages 2572–2577. IEEE, 2010.

Allison Ryan and J Karl Hedrick. Particle filter based information-theoretic active sensing. *Robotics and Autonomous Systems*, 58(5):574–584, 2010.

Victor Manuel Hernandez Bennetts, Achim J Lilienthal, Ali Abdul Khaliq, Victor Pomareda Sese, and Marco Trincavelli. Towards real-world gas distribution mapping and leak localization using a mobile robot with 3d and remote gas sensing capabilities. In *2013 IEEE International Conference on Robotics and Automation*, pages 2335–2340. IEEE, 2013.

Vijay Kumar, Daniela Rus, and Sanjiv Singh. Robot and sensor networks for first responders. *IEEE Pervasive computing*, 3(4):24–33, 2004.

Paul E Rybski, Sascha A Stoeter, Michael D Erickson, Maria Gini, Dean F Hougen, and Nikolaos Papanikolopoulos. A team of robotic agents for surveillance. In *Proceedings of the fourth international conference on autonomous agents*, pages 9–16, 2000.

Pierre-Yves Oudeyer and Frederic Kaplan. How can we define intrinsic motivation? In *the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, Lund: LUCS, Brighton, 2008.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. *Guided Self-Organization: Inception*, pages 67–114, 2014.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Josep M Porta, Nikos Vlassis, Matthijs TJ Spaan, and Pascal Poupart. Point-based value iteration for continuous pomdps. 2006.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes, we include these descriptions in the main text but put the detailed algorithm in the appendix because of the page limit.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes, we have a description of the complexity in the Sec. 6.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Yes, we submit the code as supplemental material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes, We include the assumptions in the statement of the theoretical results.
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes, we include the proofs in the appendix.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes, we include the code and instructions in the supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes, we have descriptions for the measure and error bars in the Sec. 6.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes, we have a description of the hardware settings of the machine used for the experiment at the start of Sec. 6.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Not Applicable
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not Applicable
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable

Appendix : Efficient Variational Sequential Information Control

A Global MI control objective

For convenience, we restate the MI control objective from the main text (Eqn. (2)) here:

$$\pi^* = \operatorname*{argmax}_{\pi} I(X_1^T; Y_1^T \mid \pi). \tag{32}$$

We will show that this decomposes into a sum of objectives across time, each of which depends only on a single latent state X_t, as in Eqn. (6). For simplicity of notation, we drop the dependence on the policy π. The derivation makes use of the MI chain rule Cover and Thomas (2006), namely for three random variables A, B, C the MI decomposes as:

$$I(A; \{B, C\}) = I(A; B) + I(A; C|B), \tag{33}$$

By the MI chain rule on variables Y_1^T the MI control objective in Eqn. (32) decomposes additively as:

$$\begin{aligned} I(X_1^T; Y_1^T) &= I(X_1^T; Y_1) + I(X_1^T; Y_2^T|Y_1) \\ &= I(X_1^T; Y_1) + I(X_1^T; Y_2|Y_1) + I(X_1^T; Y_3^T|Y_1) \\ &\dots \\ &= I(X_1^T; Y_1) + \sum_{t=2}^{T} I(X_1^T; Y_t|Y_1^{t-1}). \end{aligned} \tag{34}$$

The chain rule is further applied on variables X_1^T for each term in Eqn. (34). Taking the first term as an example we have,

$$I(X_1^T; Y_1) = I(X_1; Y_1) + \sum_{t=2}^{T} I(X_t; Y_1|X_1^{t-1}) = I(X_1; Y_1). \tag{35}$$

The last equality holds since $\sum_{t=2}^{T} I(X_t; Y_1|X_1^{t-1}) = 0$ because $Y_t \perp\!\!\!\perp X_{i\neq t}|X_t$ by the observation model $p(y_t \mid x_t)$. Continuing repeated application of the chain rule and the aforementioned independence each term in Eqn. (34) simplifies as,

$$I(X_1^T; Y_t \mid Y_1^{t-1}) = I(X_t; Y_t \mid Y_1^{t-1}) + I(\{X_i\}_{i\in\{1,\dots,T\}\setminus t}; Y_t \mid Y_1^{t-1}, X_t) = I(X_t; Y_t \mid Y_1^{t-1}). \tag{36}$$

Combining these steps we have the decomposed global MI objective,

$$I(X_1^T; Y_1^T) = I(X_1; Y_1) + \sum_{t=2}^{T} I(X_t; Y_t|Y_1^{t-1}). \tag{37}$$

One detail not discussed in the main text due to space limitations is that our model includes an initial state X_0, which does not appear in the MI objective Eqn. (37). There is no observation associated with this initial state X_0 so it is simply marginalized out for each $d_1 \in \mathcal{D}$ during the initial control step at $t = 1$. Explicitly incorporating the decision variable we see that the initial decision d_1 modulates the prior entropy $H(X_1 \mid d_1)$ in the first term $I(X_1; Y_1 \mid d_1) = H(X_1 \mid d_1) - H(X_1 \mid Y_1, d_1)$ and so is accounted for even when X_0 is marginalized out of the objective.

B Moment-matching in Gaussian case

We propose the Theorem 3.1 and claim that Eqn. (12) takes a closed-form solution at the moment-matching case. Due to the space limit, we defer the proof here.

B.1 Proof of theorem 3.1

Theorem 3.1 states:*Let the joint* $q(X,Y) = \mathcal{N}(m,\Sigma)$ *match the moments of any target distribution* $\hat{p}(X,Y)$. *Then the marginal* $q_m(X) = \int q(X,y)dy$ *and conditional* $q_c(X \mid Y) = q(X,Y)/q(Y)$ *minimize the upper bound Eqn.* (13).

Proof. Recall the upper bound on MI error as Eqn. (13),

$$|I_{\hat{p}_{t|t-1}} - I_{\hat{p}_{t|t-1}}(q)| \leq \min_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) + \min_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) + C. \tag{38}$$

For Gaussian marginal and conditional,

$$q_m(X_t) = \mathcal{N}(X_t \mid \mu, Q) \quad \text{and} \quad q_c(X_t \mid Y_t) = \mathcal{N}(X_t \mid AY_t + b, \Gamma), \tag{39}$$

the optimal Gaussian q_m^* that minimizes marginal cross-entropy is given by moment-matching Murphy (2012),

$$q_m^* = \operatorname*{argmin}_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) - \underbrace{H_{\hat{p}_{t|t-1}}(X_t)}_{\text{constant}} = \operatorname*{argmin}_{q_m} \mathrm{KL}(\hat{p}_{t|t-1} \,\|\, q_m), \tag{40}$$

so

$$\mu^* = E_{\hat{p}_{t|t-1}}[X_t] \quad \text{and} \quad Q^* = \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t). \tag{41}$$

For simplicity, we drop the $\hat{p}_{t|t-1}$ in the expectation and covariance calculations in the following of this proof. Now we consider the conditional objective:

$$\begin{aligned}
q_c^* &= \operatorname*{argmin}_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) \\
\text{Let } \alpha(A,b,\Gamma) &\equiv E\left[-\log \mathcal{N}(X_t \mid AY_t + b, \Gamma)\right] \\
&= \frac{1}{2}\log \mid \Gamma \mid + E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] + Const.
\end{aligned} \tag{42}$$

First, we solve for b,

$$\begin{aligned}
\nabla_b \alpha &= \nabla_b E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \\
&= E\left[\Gamma^{-1}(X_t - b - AY_t)(-\nabla_b b)\right] \\
&= -E\left[\Gamma^{-1}(X_t - b - AY_t)\right] = 0 \\
b^* &= E\left[(X_t - AY_t)\right].
\end{aligned} \tag{43}$$

Second, we solve for A,

$$\begin{aligned}
\nabla_A \alpha &= \nabla_A E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \\
&= E\left[\Gamma^{-1}(X_t - b - AY_t)(-\nabla_A AY_t)\right] \\
&= -E\left[\Gamma^{-1}(X_t - b - AY_t)Y_t^T\right] = 0 \\
0 &= E\left[X_t Y_t^T - bY_t^T - AY_t Y_t^T\right] \\
&\quad \textbf{Backsubstitute} \quad b^* \\
0 &= E\left[X_t Y_t^T - E\left[(X_t - AY_t)\right] Y_t^T - AY_t Y_t^T\right] \\
&= E\left[X_t Y_t^T\right] - E\left[X_t\right] E\left[Y_t^T\right] + \\
&\quad AE\left[Y_t\right] E\left[Y_t^T\right] - AE\left[Y_t Y_t^T\right] \\
&= \mathrm{Cov}(X_t, Y_t) - A\mathrm{Cov}(Y_t, Y_t) \\
A^* &= \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}
\end{aligned} \tag{44}$$

Back substitute into b^*,

$$b^* = E\left[X_t - AY_t\right] = E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1} E\left[Y_t\right]. \tag{45}$$

The full conditional mean is then,

$$\begin{aligned} A^*Y_t + b^* &= \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}Y_t + E\left[X_t\right] - \\ &\qquad \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}E\left[Y_t\right] \\ &= E\left[X_t\right] + \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right]) \end{aligned} \tag{46}$$

Solve for conditional covariance Γ,

$$\begin{aligned} \nabla_\Gamma\, \alpha &= \nabla_\Gamma \frac{1}{2}\log \mid \Gamma \mid + \frac{1}{2}E\left[tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \\ &= \Gamma - E\left[(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] = 0 \\ \Gamma^* &= E\left[(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \end{aligned} \tag{47}$$

Back substitute $A^*Y_t + b^*$,

$$\begin{aligned} \Gamma^* &= E[(X_t - E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right]) \\ &\quad (X_t - E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right])^T] \\ &= E[(X_t - E\left[X_t\right])(X_t - E\left[X_t\right])^T] \\ &\quad - 2E[(X_t - E\left[X_t\right])(Y_t - E\left[Y_t\right])^T]\mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1} \\ &\quad + \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}E[(Y_t - E\left[Y_t\right])(Y_t - E\left[Y_t\right])^T] \\ &\quad * \mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(X_t, Y_t)^T \\ &= \mathrm{Cov}(X_t, X_t) - 2\mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(X_t, Y_t)^T \\ &\quad + \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(Y_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(X_t, Y_t)^T \\ &= \mathrm{Cov}(X_t, X_t) - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(X_t, Y_t)^T. \end{aligned}$$

Now we consider $q^{(d_t)}_{t|t-1}(X_t, Y_t) = \mathcal{N}(m, \Sigma)$ moment-matched to $\hat{p}_{t|t-1}(X_t, Y_t)$,

$$\begin{aligned} m &= \begin{bmatrix} m_X \\ m_Y \end{bmatrix} = \begin{bmatrix} E_{\hat{p}_{t|t-1}}[X_t] \\ E_{\hat{p}_{t|t-1}}[Y_t] \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, X_t) & \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, Y_t) \\ \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, Y_t)^T & \mathrm{Cov}_{\hat{p}_{t|t-1}}(Y_t, Y_t) \end{bmatrix} \end{aligned} \tag{48}$$

Thus,

$$q_m = \mathcal{N}(m_X, \Sigma_X) = \mathcal{N}(\mu^*, Q^*) = \operatorname*{argmin}_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)). \tag{49}$$

The conditional Gaussian from the joint $q^{(d_t)}_{t|t-1}$ are,

$$q_c = \mathcal{N}(X_t \mid \underbrace{m_X + \Sigma_{XY}\Sigma_Y^{-1}(Y_t - m_Y)}_{A^*Y_t + b^*}, \underbrace{\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T}_{\Gamma^*}) = \operatorname*{argmin}_{q_c} H_{\hat{p}_{t|t-1}}(q_c). \tag{50}$$

Therefore, moment-matching the augmented distribution to a joint Gaussian distribution yields optimal Gaussian marginal and conditional approximations that minimize the MI error bound, when q_m and q_c are considered as Gaussian distributions. In other words, optimal q_m^* and q_c^* share the same joint distribution when q_m and q_c are considered as Gaussian distributions. □

According to the Foster et al. Foster et al. (2019), the MI error bound can be restated as,

$$|I_{\hat{p}_{t|t-1}} - I_{\hat{p}_{t|t-1}}(q)| \leq \min_{q_m} H_{\hat{p}_{t|t-1}}(q_{ml}(Y_t)) + \min_{q_c} H_{\hat{p}_{t|t-1}}(q_{cl}(Y_t \mid X_t)) + C, \tag{51}$$

where C does not depend on q_{ml} or q_{cl}. So we could have a similar corollary as Theorem 3.1.

Corollary B.1. *When q_{ml} and q_{cl} are Gaussian distributions, moment-matching $q^{(d_t)}_{t|t-1}(X_t, Y_t) = \mathcal{N}(m, \Sigma)$ to $\hat{p}_{t|t-1}(X_t, Y_t)$, will yield optimal $q^*_{ml} = \int q^{(d_t)}_{t|t-1}(x_t, Y_t)\mathrm{d}x_t$ and $q^*_{cl} = \frac{q^{(d_t)}_{t|t-1}(X_t, Y_t)}{\int q^{(d_t)}_{t|t-1}(X_t, y_t)\mathrm{d}y_t}$.*

The proof is identical to that of Theorem 3.1 but swap X_t and Y_t. Therefore, we don't reiterate here.

B.2 Closed-form solution for MI

Lemma B.2. *Given* $q^{(d_t)}_{t|t-1}(X_t, Y_t) = \mathcal{N}(m, \Sigma)$, *which moment-matches* $\hat{p}_{t|t-1}(X_t, Y_t)$, $I_{\hat{p}_{t|t-1}}(q) \equiv H_{\hat{p}_{t|t-1}}(q_m(X_t)) - H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t))$ *has a closed-form solution, where* q_m *and* q_c *are marginal and conditional distribution of* $q^{(d_t)}_{t|t-1}$ *respectively.*

Proof. Given that q_m and q_c are Gaussian distributions, we have proved in theorem 3.1 that they share the same joint distribution. Thus, we have

$$I_{\hat{p}_{t|t-1}}(q) = H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) - H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t)). \tag{52}$$

Since q_m and q_c are both Gaussian distributions, w.l.o.g., we prove that $H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t))$ has a closed-form solution, and it applies to $H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t))$. Assume $q_{t|t-1}(X_t) = \mathcal{N}(X_t \mid m_X, \Sigma_X)$ and the dimension of Σ_X is k,

$$\begin{aligned} H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) &= -E_{\hat{p}_{t|t-1}}[\log \mathcal{N}(X_t \mid m_X, \Sigma_X)] \\ &= -E_{\hat{p}_{t|t-1}}[-\frac{k}{2}\log 2\pi - \frac{1}{2}\log \mid \Sigma_X \mid -\frac{1}{2}(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] \\ &= \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X \mid + \frac{1}{2} E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] \end{aligned} \tag{53}$$

We assume that Σ_X is a valid covariance matrix, thus it could be decomposed as $\Sigma_X = LL^T$.

$$E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] = E_{\hat{p}_{t|t-1}}[(L^{-1}X_t - L^{-1}m_X)^T (L^{-1}X_t - L^{-1}m_X)] \tag{54}$$

Let $C \equiv E_{\hat{p}_{t|t-1}}[(L^{-1}X_t - L^{-1}m_X)(L^{-1}X_t - L^{-1}m_X)^T]$, and $E_{\hat{p}_{t|t-1}}[L^{-1}X_t] = L^{-1}m_X$. By definition,

$$\begin{aligned} C &= \mathrm{Cov}(L^{-1}X_t, L^{-1}X_t) \\ &= L^{-1} \underbrace{\mathrm{Cov}(X_t, X_t)}_{\Sigma_X} L^{-T} \\ &= L^{-1} L L^T L^{-T} \\ &= I. \end{aligned}$$

Therefore,

$$E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] = tr(C) = k. \tag{55}$$

Summing terms up, we have a closed-form solution for

$$H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X \mid + \frac{k}{2}. \tag{56}$$

A similar result applies to $H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t))$, but replace Σ_X with $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T$

$$H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t)) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \mid + \frac{k}{2}. \tag{57}$$

$$I_{\hat{p}_{t|t-1}}(q) = \frac{1}{2}\log \mid \Sigma_X \mid - \frac{1}{2}\log \mid \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \mid . \tag{58}$$

□

Theorem 3.2 states: Let the joint $q(X, Y) = \mathcal{N}(m, \Sigma)$ match the moments of any target distribution $\hat{p}(X, Y)$, then we have that $H_{\hat{p}}(q) = H_q(q)$.

Proof. We have proved as Eqn. (56), when q is a Gaussian distribution with k-dimension covariance matrix Σ_q that moment-matches $\hat{p}$, the cross entropy is

$$H_{\hat{p}}(q) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_q \mid + \frac{k}{2}. \tag{59}$$

By definition, the entropy of Gaussian distribution q is

$$H_q(q) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_q \mid + \frac{k}{2} = H_{\hat{p}}(q). \tag{60}$$

□

C Algorithm statement of sequential variational information control

Algorithm 1 provides a complete statement of our proposed method for sequential variational MI control in the greedy setting. The algorithm is provided for, both, EP and ADF inference with relevant components to each denoted by color. ADF is a special case of EP, consisting of only the first forward pass of inference, and is denoted in (red). Additional (blue) lines are specific to EP as the iterate forward-and-backward message updates. The MI approximation and decision selection are equivalent for both cases.

Algorithm 1 Sequential Variational Information Control

Input: Start state x_0,prior distribution $x_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$
Output: A series of decisions $D_{1:T}$
Initialization: $\alpha_0(x_0) = \mathcal{N}(\mu_0, \Sigma_0), \beta_{0:T-1} = 1$
Let $D[t]$ be the optimal decision at time t when the final step is t,$1 \le t \le T$
{// Estimate MI for each decision at time t}
for $t = 1$ **to** T **do**
  for $d_t = 1$ **to** K **do**
    $\hat{p}_{t,d_t}(X_t, Y_t) = \int \alpha_{t-1}(x_{t-1}) p(X_t|x_{t-1}, d_t) p(Y_t|X_t)\, dx_{t-1}$ {// Augmented predictive distribution at time t}
    $q_{t,d_t}(X_t, Y_t) = \text{argmin}_q\ KL(q||\hat{p}_{t,d_t})$ {// KL-projection}
  end for
  $d_t^* = \text{argmax}_{d_t \in \{1,\dots,K\}}\ I_{\hat{p}_{t,d_t}}(q_{t,d_t})$ {// Choose decision with maximum MI per Sec. 3.2 or Sec. 4.2}
  Execute decision d_t^* and observe $Y_t = y_t$
  {// ADF-update: (always do this)}
  $\hat{p}_t(X_t) \propto \int \alpha_{t-1}(x_{t-1}) p(X_t \mid x_{t-1}, d_t^*) p(y_t|X_t) d\, x_{t-1}$ {// Augmented filter distribution at time t}
  $\alpha_t(X_t) = \text{argmin}_\alpha\ \text{KL}(\hat{p}_t \,\|\, \alpha)$ {// KL-projection–forward message update}
  {// EP-update: (only if doing EP)}
  repeat
    for $i = 1$ **to** t **do**
      $\hat{p}_i(X_i) = \int \alpha_{i-1}(x_{i-1}) p(X_i \mid x_{i-1}, d_i^*) p(y_i \mid X_i) \beta_i(X_i)\, dx_{i-1}$
      $q_i(X_i) = \text{argmin}_q\ \text{KL}(\hat{p}_i \,\|\, q)$
      $\alpha_i(X_i) \propto \frac{q_i(X_i)}{\beta_i(X_i)}$ {// EP Forward message update}
    end for
    for $i = (t-1)$**to** 0 **do**
      $\hat{p}(X_i) = \int \alpha_i(X_i) p(x_{i+1} \mid x_i, d_{i+1}^*) p(y_{i+1} \mid x_{i+1}, d_{i+1}^*) \beta_{i+1}(x_{i+1})\, dx_{i+1}$
      $q_i(X_i) = \text{argmin}_q\ \text{KL}(\hat{p}_i \,\|\, q)$
      $\beta_i(X_i) \propto \frac{q_i(X_i)}{\alpha_i(X_i)}$ {// EP Backward message update}
    end for
  until $\{\alpha_i, \beta_i\}$ converge
end for

D Theorem 4.1

Theorem 4.1 states, *In a model with linear-Gaussian observations, given the augmented distribution at time t,* $\hat{p}_{t|t-1}(X_t, Y_t \mid \mathcal{H}_{t-1}, d_t)$ *and its Gaussian approximation,* $q_{t|t-1}^{(d_t)}(X_t, Y_t)$*, by moment-matching,*

$$I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid \mathcal{H}_{t-1}, d_t) \le I_{q_{t|t-1}^{(d_t)}}(X_t; Y_t \mid \mathcal{H}_{t-1}, d_t)$$

Proof. For brevity, we drop the time and decision index for $\hat{p}_{t|t-1}$ and $q_{t|t-1}^{(d_t)}$ in the proof below. According to the definition of the information control setting and mutual information, the mutual information between X_t and Y_t for the augmented distribution $\hat{p}$ can be decomposed as

$$I_{\hat{p}}(X_t; Y_t \mid d_t) = H_{\hat{p}}(Y_t \mid d_t) - H_{\hat{p}}(Y_t \mid X_t) \tag{61}$$

Similarly, the mutual information for the approximation q is decomposed as

$$I_q(X_t; Y_t) = H_q(Y_t \mid d_t) - H_q(Y_t \mid X_t). \tag{62}$$

Since Gaussian $\hat{p}(X_t, Y_t) \approx q(X_t, Y_t)$ by moment-matching, and according to Theorem 3.2,

$$H_q(Y_t \mid d_t) = H_p(q(Y_t \mid d_t)). \tag{63}$$

With Lemma B.2, the MI $I_q(X_t; Y_t)$ can be solved in closed-form. Moreover, by Corollary B.1,

$$I_{\hat{p}}(q) = H_{\hat{p}}(q(Y_t \mid d_t)) - H_{\hat{p}}(q(Y_t \mid X_t)) \tag{64}$$

optimizes the MI error bound as Eqn. 51. Specially, when $\hat{p}(Y_t \mid X_t) = \mathcal{N}(Y_t \mid HX_t + b, R)$ (R is k-dimensional), the entropy is

$$H_{\hat{p}}(Y_t \mid X_t) = \frac{k}{2} \log 2\pi + \frac{1}{2} \log \mid R \mid + \frac{k}{2} \tag{65}$$

by definition. Since Gaussian distribution $q(Y_t \mid X_t)$ moment-matches $\hat{p}(Y_t \mid X_t)$, according to Theorem 3.2,

$$H_{\hat{p}}(Y_t \mid X_t) = H_q(Y_t \mid X_t) = \frac{k}{2} \log 2\pi + \frac{1}{2} \log \mid R \mid + \frac{k}{2}. \tag{66}$$

So

$$I_q(X_t; Y_t \mid d_t) = H_q(Y_t \mid d_t) - H_q(Y_t \mid X_t) = H_q(Y_t \mid d_t) - H_{\hat{p}}(Y_t \mid X_t) \tag{67}$$

On the other hand, by the Gibbs inequality,

$$H_{\hat{p}}(Y_t \mid d_t) \leq H_{\hat{p}}(q(Y_t \mid d_t)) = H_q(Y_t \mid d_t). \tag{68}$$

Thus,

$$\begin{aligned} H_{\hat{p}}(Y_t \mid d_t) - H_{\hat{p}}(Y_t \mid X_t) \leq H_q(Y_t \mid d_t) - H_{\hat{p}}(Y_t \mid X_t) = H_q(Y_t \mid d_t) - H_q(Y_t \mid X_t) \\ I_{\hat{p}}(X_t; Y_t \mid d_t) \leq I_q(X_t; Y_t \mid d_t) \end{aligned} \tag{69}$$

□

E Constrained GMM approximation

E.1 Moment-matching calculation

At time t, let the approximated filter distribution $q(X_t|y_1^t, d_1^t) = \sum_{i=1}^K \pi_i \mathcal{N}(X_t|\mu_i, \Sigma_i)$, given the GMM-Gaussian system, i.e.,

$$p_d(X_{t+1}|X_t) = \sum_{i=1}^N w_d(i) \mathcal{N}(X_{t+1}|A_{d,i}X_i, C_{d,i}), \tag{70}$$

$$p(Y_{t+1}|X_{t+1}) = \mathcal{N}(Y_{t+1} \mid HX_{t+1}, R), \tag{71}$$

for a given decision d

$$\begin{aligned} \hat{p}(X_{t+1}, Y_{t+1}, S_{t+1} = j) &= \int \sum_{i=1}^K \pi_i \mathcal{N}(X_t|\mu_i, \Sigma_i) w_j \mathcal{N}(X_{t+1}|A_j X_t, C_j) p(Y_{t+1}|HX_{t+1}, R) d_{X_t} \\ &= \sum_{i=1}^K \pi_i w_j \mathcal{N}(Y_{t+1}|HX_{t+1}, R) \mathcal{N}(X_{t+1} \mid A_j\mu_i, C_j + A_j\Sigma_i A_j^T) \\ &= \sum_{i=1}^K \pi_i w_j \mathcal{N}(Y_{t+1}|HX_{t+1}, R) \mathcal{N}(X_{t+1} \mid m_{ij}, P_{ij}) \end{aligned} \tag{72}$$

where we define $P_{ij} = C_j + A_j\Sigma_i A_j^T$, $m_{ij} = A_j\mu_i$.
Applying constrained GMM projection to approximate $p(X_{t+1}, Y_{t+1}|S_{t+1} = j)$ by

$$q(X_{t+1}, Y_{t+1}, S_{t+1} = j) = \mathcal{N}(Y_{t+1}|\eta, P) w_j \mathcal{N}(X_{t+1}|F_j Y_{t+1}, M_j). \tag{73}$$

We could first compute P as the projection of $\text{Cov}_{\hat{p}}(Y)$–the covariance of Y under the augmented distribution $\hat{p}(Y)$:

$$\begin{aligned}\hat{p}(Y_{t+1}) &= \sum_{i=1}^{K} \pi_i \sum_{j=1}^{N} w_j \int \mathcal{N}(Y_{t+1} \mid Hx_{t+1}, R)\mathcal{N}(x_{t+1}|m_{ij}, P_{ij})dx_{t+1} \\ &= \sum_{i=1}^{K} \pi_i \sum_{j=1}^{N} w_j \mathcal{N}(Y_{t+1} \mid Hm_{ij}, R + HP_{ij}H^T) \end{aligned} \tag{74}$$

Let

$$\begin{aligned} M &= \sum_{i=1}^{K} \pi_i \sum_{j=1}^{N} w_j (Hm_{ij}), \\ V &= \sum_{i=1}^{K} \pi_i \sum_{j=1}^{N} w_j [R + HP_{ij}H^T + (Hm_{ij})(Hm_{ij})^T] - MM^T \end{aligned} \tag{75}$$

$$\begin{aligned} P &= \text{Cov}_{\hat{p}}(Y) = V \\ \eta &= M \end{aligned} \tag{76}$$

To compute F_j and M_j, we consider

$$\begin{aligned} G(F_j, M_j) &\equiv \min_{F_j, M_j} E\left[-\log \mathcal{N}(X_t \mid F_j Y_t, M_j)\right] \\ &= \min_{F_j, M_j} \frac{1}{2} \log \mid M_j \mid + E\left[\frac{1}{2} tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right]. \end{aligned} \tag{77}$$

Solve for F_j,

$$\begin{aligned} \nabla_{F_j} G &= \nabla_{F_j} E\left[\frac{1}{2} tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\ &= E\left[M_j^{-1}(X_t - F_j Y_t)(-\nabla_{F_j} F_j Y_t)\right] \\ &= -E\left[M_j^{-1}(X_t - F_j Y_t) Y_t^T\right] = 0 \\ 0 &= E\left[X_t Y_t^T - F_j Y_t Y_t^T\right] \\ F_j^* &= E\left[X_t Y_t^T\right] E\left[Y_t Y_t^T\right]^{-1} \\ &= \{\text{Cov}(X_t, Y_t) + E[X_t]E[Y_t^T]\}\{\text{Cov}(Y_t, Y_t) + E[Y_t]E[Y_t^T]\}^{-1} \end{aligned} \tag{78}$$

Solve for M_j,

$$\begin{aligned} \nabla_{M_j} G &= \nabla_{M_j} \frac{1}{2} \log \mid M_j \mid + \frac{1}{2} E\left[tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\ &= M_j - E\left[(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] = 0 \\ M_j^* &= E\left[(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\ &= \text{Cov}(X_t, X_t) + E[X_t]E[X_t^T] - \{\text{Cov}(X_t, Y_t) + E[X_t]E[Y_t^T]\} \\ &\quad \{\text{Cov}(Y_t, Y_t) + E[Y_t]E[Y_t^T]\}^{-1}\{\text{Cov}(Y_t, X_t) + E[Y_t]E[X_t^T]\} \end{aligned} \tag{79}$$

Moreover, we project $\text{Cov}_{\hat{p}}(X_{t+1}|S_{t+1} = j)$ to $\text{Cov}_q(X_{t+1}|S_{t+1} = j)$ for MI estimation, which is shown later.

$$\hat{p}(X_{t+1}, Y_{t+1}, S_{t+1} = j) = \Sigma_{i=1}^{K} \pi_i w_j \mathcal{N}(Y_{t+1}|HX_{t+1}, R)\mathcal{N}(X_{t+1}|m_{ij}, P_{ij}) \tag{80}$$

$$\begin{aligned} \hat{p}(X_{t+1}, Y_{t+1}|S_{t+1} = j) &= \Sigma_{i=1}^{K} \pi_i \mathcal{N}(Y_{t+1}|HX_{t+1}, R)\mathcal{N}(X_{t+1}|m_{ij}, P_{ij}) \\ &= \Sigma_{i=1}^{K} \pi_i N\left(\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} \middle| \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix}, \begin{bmatrix} P_{ij} & P_{ij}H^T \\ HP_{ij} & R + HP_{ij}H^T \end{bmatrix}\right) \end{aligned} \tag{81}$$

Let

$$\widetilde{\mu}_{t+1} = \Sigma_{i=1}^{K} \pi_i \begin{bmatrix} m_{ij} \\ H m_{ij} \end{bmatrix},$$
$$\widetilde{V}_{t+1} = \Sigma_{i=1}^{K} \pi_i \left\{ \begin{bmatrix} P_{ij} & P_{ij}H^T \\ HP_{ij} & R + HP_{ij}H^T \end{bmatrix} + \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix} \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix}^T \right\}$$
$$-\widetilde{\mu}_{t+1}\widetilde{\mu}_{t+1}^T \tag{82}$$

Let

$$m_j = \Sigma_{i=1}^{K} \pi_i m_{ij}, V_j = \Sigma_{i=1}^{K} \pi_i [P_{ij} + m_{ij}m_{ij}^T] - m_j m_j^T, \tag{83}$$

then

$$\mathrm{Cov}_{\hat{p}}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} V_j & V_j H^T \\ HV_j & R + HV_j H^T \end{bmatrix}, \tag{84}$$

$$\mathrm{Cov}_{\hat{p}}^{-1}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} V_j^{-1} + H^T R^{-1} H & -H^T R^{-1} \\ -R^{-1}H & R^{-1} \end{bmatrix}, \tag{85}$$

$$\mathrm{Cov}_{\hat{p}}(X_{t+1}|S_{t+1} = j) = V_j. \tag{86}$$

Since

$$\mathrm{Cov}_{q}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} M_j + F_j P F_j^T & F_j P \\ P F_j^T & P \end{bmatrix} \tag{87}$$

and

$$\mathrm{Cov}_{q^{-1}}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} M_j^{-1} & -M_j^{-1}F_j \\ -F_j^T M_j^{-1} & P^{-1} + F_j^T M_j^{-1} F_j \end{bmatrix}, \tag{88}$$

$$\mathrm{Cov}_{q}(X_{t+1}|S_{t+1} = j) = M_j + F_j P F_j^T = V_j, \tag{89}$$

and

$$\mathrm{Cov}_{q}(X_{t+1}|Y_{t+1}, S_{t+1} = j) = M_j. \tag{90}$$

E.2 MI of Constrained GMM approximation

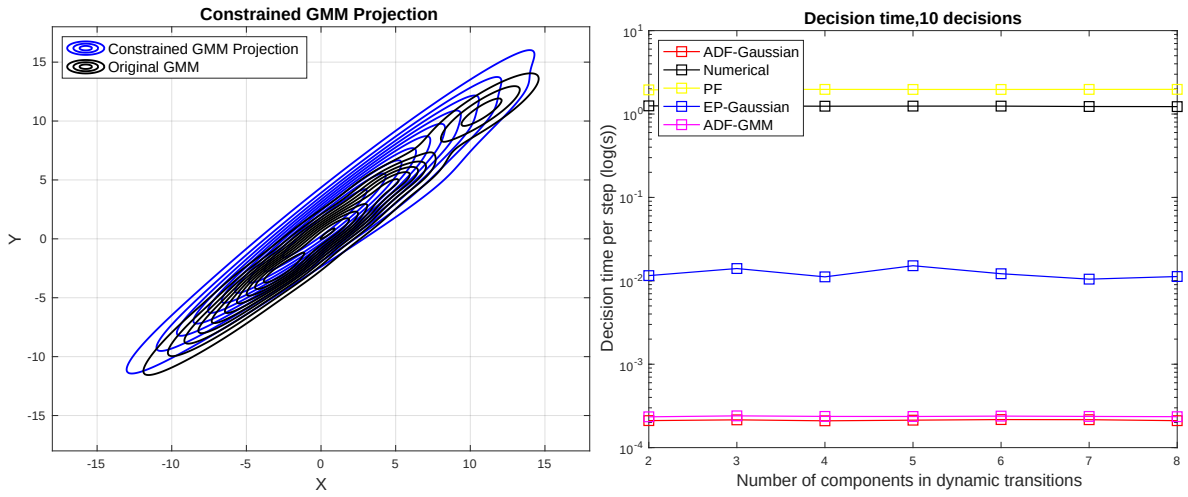
Our goal is to compute

$$\begin{aligned} I_q(\{X_{t+1}, S_{t+1}\}; Y_{t+1}) &= H_q(X_{t+1}, S_{t+1}) - H_q(X_{t+1}, S_{t+1}|Y_{t+1}) \\ &= H_q(X_{t+1}|S_{t+1}) + H_q(S_{t+1}) \\ &\quad - H_q(X_{t+1}|S_{t+1}, Y_{t+1}) - H_q(S_{t+1}|Y_{t+1}). \end{aligned} \tag{91}$$

Due to the dependence shown as the Fig. 2, $H(S_{t+1}|Y_{t+1}) = H(S_{t+1})$. We can further simplify Eqn. (91) by

$$\begin{aligned} I_q(\{X_{t+1}, S_{t+1}\}; Y_{t+1}) &= H(X_{t+1}|S_{t+1}) - H(X_{t+1}|S_{t+1}, Y_{t+1}) \\ &= \sum_{j}^{N} q(S_{t+1} = j)[H(X_{t+1}|S_{t+1} = j)] \\ &\quad - \sum_{j}^{N} q(S_{t+1} = j)[H(X_{t+1}|S_{t+1} = j, Y_{t+1})]. \end{aligned} \tag{92}$$

Because both $q(X_{t+1}|S_{t+1} = j)$ and $q(X_{t+1}|S_{t+1} = j, Y_{t+1})$ are Gaussian distribution, their entropy only involves their covariance matrix respectively. And we have computed $\mathrm{Cov}_q(X_{t+1}|S_{t+1} = j)$ as Eqn. (89) and $\mathrm{Cov}_q(X_{t+1}|S_{t+1} = j, Y_{t+1})$ as Eqn. (90). Moreover, $q(S_{t+1}) = p(S_{t+1})$. Thus, Eqn. (92) can be analytically



(a) Constrained GMM projection (b) Runtime versus components

Figure 7: (a) This figure shows that our methods perform well in approximating the $K^2 = 16$-component Gaussian mixture model. The constrained GMM projection method, represented in blue, performs a constrained projection to a $K = 4$-component GMM. (b) We fix the number of decisions to 10 and compare the average running time as the number of components in the GMM dynamic transition grows from 2 to 8. We can see that our approaches perform consistently as the number of components grows.

calculated as

$$\begin{aligned} I_q(\{X_{t+1}, S_{t+1}\}; Y_{t+1}) &= H_q(X_{t+1}, S_{t+1}) - H_q(X_{t+1}, S_{t+1}|Y_{t+1}) \\ &= H_q(X_{t+1}|S_{t+1}) - H_q(X_{t+1}|S_{t+1}, Y_{t+1}) \\ &= \frac{1}{2}\sum_j^N w_j \log|2\pi e(M_j + F_j P F_j^T)| - \frac{1}{2}\sum_j^N w_j \log|2\pi e M_j| \end{aligned} \tag{93}$$

F Additional experimental results

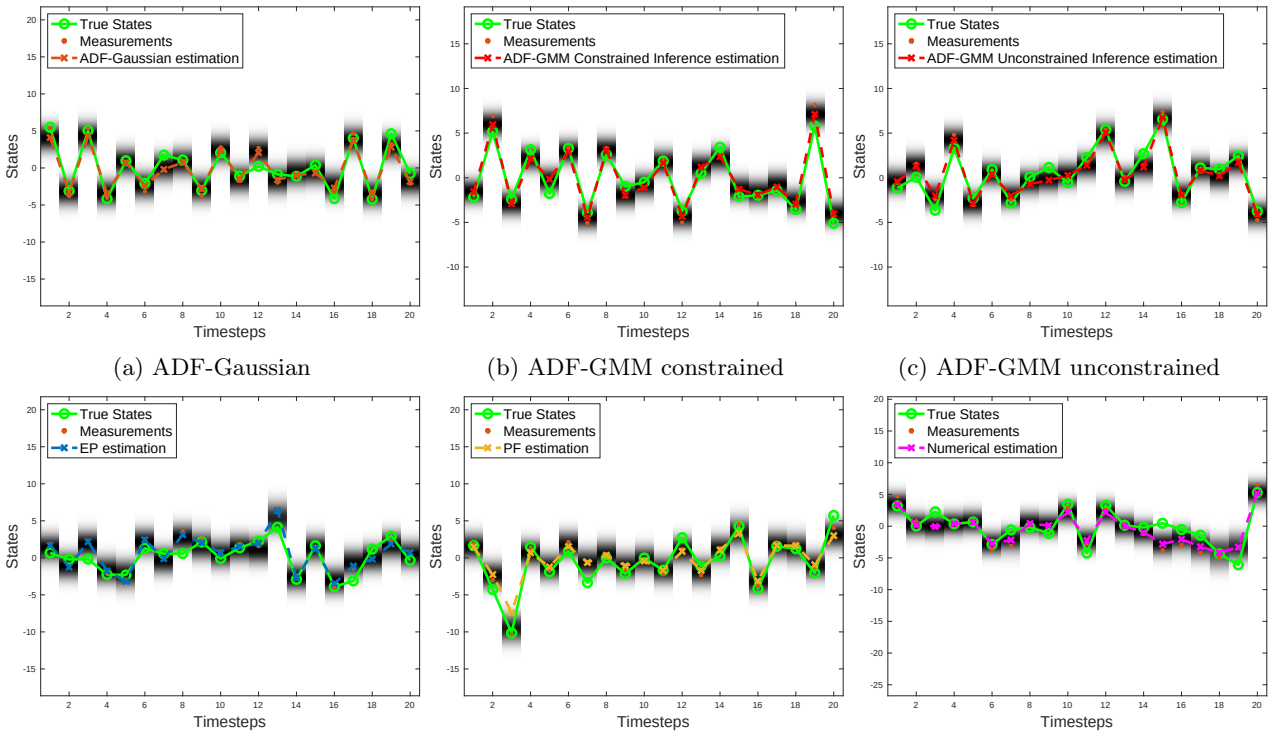
Due to space limitations in the main text, we provide only a limited set of experimental results for variational MI control. We provide a more detailed and extensive empirical evaluation here, considering a more comprehensive array of configurations than in the main text.

F.1 Single-step scenario

We validate that the constrained GMM projection has good performance in approximating the target distribution. The approximation result of one example is shown in Fig. 7a. The constrained GMM projection projects the true distribution down to a K-component Gaussian mixture. A plot of the PGM of constrained GMM approximation is shown as Fig. 2.

F.2 GMM-Gaussian Control

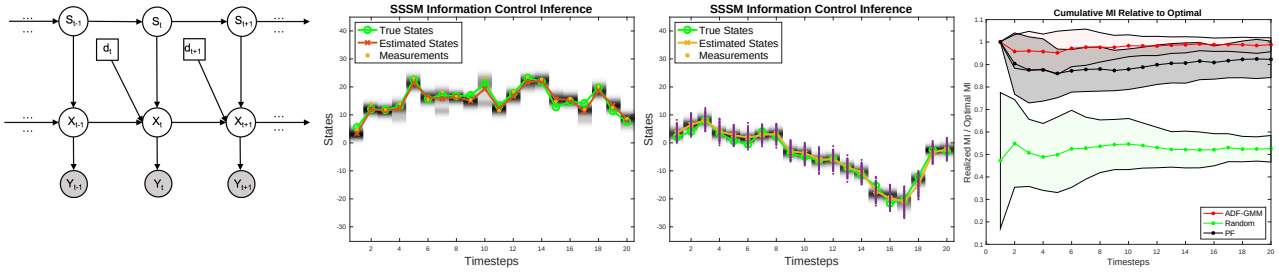
Fig. 8a to Fig. 8f show qualitative comparisons of inference for individual runs of each method. Grayscale bars in the plots represent the filter distribution approximated by the numerical approximation method. The lines are the estimations of true states by each method respectively. Apart from that, we also show our methods perform closely in posterior inference to the numerical method in the setting of the same trajectory, i.e., all the methods share the same decision path and observed measurements. The result is shown in Fig. 3b. We have to reiterate that our method is much faster than PF or the numerical approximation when the sampling size/number of bins is large (e.g., 3000) for both methods. Another speedup test is shown in Fig. 7b.



(a) ADF-Gaussian (b) ADF-GMM constrained (c) ADF-GMM unconstrained

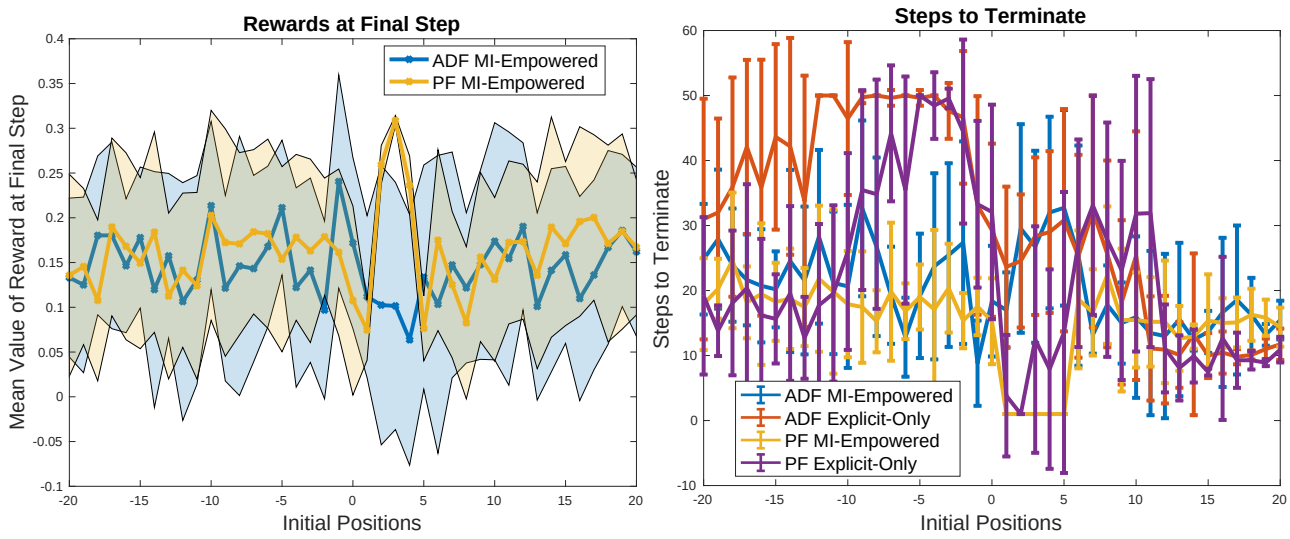
(d) EP-Gaussian (e) PF (f) Numerical

Figure 8: These figures display the posterior inference of each method separately. Different methods would have a different decision trajectory, upon which with the corresponding measurements, the 'exact inference' is approximated by the numerical approximation. And the 'exact' posterior distribution density is shown by the grayscale blocks. The ADF-Gaussian and EP-Gaussian project the target distribution to a single Gaussian distribution, and the ADF-GMM(constrained/unconstrained) maintains a GMM approximation of the filter distribution. ADF-GMM constrained method performs the ADF update using the projected distribution in estimating the MI, while the ADF-GMM unconstrained method drops the projected distribution and applies an unconstrained projection for inference.



(a) **PGM** (b) ADF-GMM inference (c) PF inference (d) Information gained

Figure 9: (a) This figure shows the **PGM** of a SSSM: a dynamic hybrid latent continuous state X_t and discrete state S_t with observations Y_t. Control inputs $d_t \in \mathcal{D}$ (decisions) parameterize the transition distribution $p(x_t \mid x_{t-1}, s_t = j, d_t)$.(b) This figure is the posterior inference of the state from the ADF-GMM. The 'exact inference' shown in grayscale blocks, is computed by the numerical approximation using the trajectory and measurement generated by the ADF-GMM method. The ADF-GMM maintains a K-component Gaussian filter distribution, so the estimation of states is computed by the mean of the K-component Gaussian posterior. (c) This figure shows the inference results of the PF with the particles it sampled through the decision-making process. (d) We show that our method outperforms the PF in terms of the decision quality, measured by the ratio of MI acquired against the optimal MI.



(a) The last exact explicit reward

(b) Steps to terminate

Figure 10: (a) This figure shows the evaluations of the exact explicit reward at the last step. Our approach is as accurate as PF with MI-empowered reward. (b) This figure shows the advantage of our method from another perspective, which says it reaches the correct door as fast as PF does. Also, we observe that our method with MI-empowered reward appears more stable than those without MI rewards, i.e., it reaches the goal using almost the same time steps from any initial position. On the other hand, as we have shown in Sec. 6.1, our approach is more computationally efficient when the number of samples is large enough to approximate the truth.

F.3 Switching state space model

The switching state-space model (SSSM) is a natural extension of the GMM-Gaussian model in the sense that marginalizing discrete switching states yield mixture dynamics. The SSSM is a more expressive model as dynamics mixtures are not constant over time. Marginalizing the switching states corresponds to mixture dynamics with an exponential number of components at each time, due to Markov dependence on the discrete switching states Murphy (2012). As shown in Fig. 9a the consists of continuous latent states $\{X_t\}_{t=1}^T \in \mathcal{X}$, discrete latent states $\{S_t\}_{t=1}^T \in \mathcal{S}$ and continuous measurements $\{Y_t\}_{t=1}^T \in \mathcal{Y}$. Due to space limitations, we defer the whole experiment here. For the SSSM we focus our comparison on three methods: ADF-GMM, Rao-Blackwellized PF, and random guess (as a control). In all settings, we also compute numerical inference and optimal greedy decisions via Reimann sum approximation.

The SSSM generative model is given by Markov transitions on the discrete states: $p(S_t = k|S_{t-1} = i) = \rho_{ki}$. Continuous state transitions and observations are then given by the remaining components of the forward model:

$$p(S_t|X_{t-1}, S_t = k, d_t) = \mathcal{N}(X_t|A_{k,d_t}X_{t-1}, Q_{k,d_t}), \quad p(Y_t|X_t) = \mathcal{N}(Y_t|HX_t, R) \tag{94}$$

We perform 11 individual trials and find that quantitative and qualitative results are largely consistent with the GMM-Gaussian experiment in Sec. 6.1. Fig. 9b shows qualitative inference results of ADF-GMM. In general, we observe comparable qualitative inferences between each of these methods (ADF and Rao-Blackwellized PF), Fig. 9b and Fig. 9c. Similarly, Fig. 9d reports mean +/- STDEV cumulative MI using the same evaluation methodology as in the GMM-Gaussian example of Sec. 6.1. It shows that our method produces better decisions in terms of the MI acquired in this model. Both control methods outperform the random guess baseline.

F.4 MI-empowered C-POMDP

We re-implemented the C-POMDP environment with a discrete decision set in Porta et al. (2006). The detailed environment setup is shown as follows. Along with the environment, we also demonstrate the mathematical derivation of posterior inference using ADF and expected reward estimation. We also attach two extra experimental results in Fig. 10a and Fig. 10b to show the advantages of MI-empowered reward and our method.

State Dynamic Transition:

$$p(X_{t+1} \mid X_t) = \begin{cases} \mathcal{N}(X_t - 2, 0.05), & \text{if } d = 0, \text{ move to the left.} \\ \mathcal{N}(X_t + 2, 0.05), & \text{if } d = 1, \text{ move to the right.} \\ \mathcal{N}(X_t, 0.05), & \text{if } d = 2, \text{ enter the door.} \end{cases} \tag{95}$$

Measurement Model:

$$p(Y_t \mid X_t) \propto \sum_{i=1}^{n} \mathcal{N}^{-1}(X_t \mid \eta_i, \Lambda_i)\mathcal{N}(Y_t \mid \mu_i, \Sigma_i), \tag{96}$$

where η_i and Λ_i are natural parameters of the Gaussian distribution. We define the Gaussian distribution in this format to facilitate further calculation concerning the Gaussian distribution. Given $\hat{p}(X_t) = \sum_{k=1}^{K} w_k \mathcal{N}(X_t \mid m_k, P_k)$ and a control selected,

$$\hat{p}(X_{t+1}) = \begin{cases} \sum_{k=1}^{K} w_k \mathcal{N}(m_k - 2, P_k + 0.05) = \sum_{k=1}^{K} w_k \mathcal{N}^{-1}(\frac{m_k - 2}{P_k + 0.05}, \frac{1}{P_k + 0.05}), & \text{if } d = 0. \\ \sum_{k=1}^{K} w_k \mathcal{N}(m_k + 2, P_k + 0.05) = \sum_{k=1}^{K} w_k \mathcal{N}^{-1}(\frac{m_k + 2}{P_k + 0.05}, \frac{1}{P_k + 0.05}), & \text{if } d = 1. \\ \sum_{k=1}^{K} w_k \mathcal{N}(m_k, P_k + 0.05) = \sum_{k=1}^{K} w_k \mathcal{N}^{-1}(\frac{m_k}{P_k + 0.05}, \frac{1}{P_k + 0.05}), & \text{if } d = 2. \end{cases} \tag{97}$$

Define

$$\phi_{ki,d} = \begin{cases} \mathcal{N}(\frac{\eta_i}{\Lambda_i} \mid m_k - 2, \frac{1}{\Lambda_i} + P_k + 0.05) & \text{if } d = 0. \\ \mathcal{N}(\frac{\eta_i}{\Lambda_i} \mid m_k + 2, \frac{1}{\Lambda_i} + P_k + 0.05) & \text{if } d = 1. \\ \mathcal{N}(\frac{\eta_i}{\Lambda_i} \mid m_k, \frac{1}{\Lambda_i} + P_k + 0.05) & \text{if } d = 2. \end{cases} \tag{98}$$

the augmented distribution is,

$$\hat{p}(X_{t+1}, Y_{t+1}) \propto \begin{cases} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k * \phi_{ki,0} \mathcal{N}^{-1}(X_{t+1} \mid \eta_i + \frac{m_k - 2}{P_k + 0.05}, \Lambda_i + \frac{1}{P_k + 0.05}) \\ \qquad \mathcal{N}(Y_{t+1} \mid \mu_i, \Sigma_i), & \text{if } d = 0. \\ \sum_{i=1}^{n} \sum_{k=1}^{K} w_k * \phi_{ki,1} \mathcal{N}^{-1}(X_{t+1} \mid \eta_i + \frac{m_k + 2}{P_k + 0.05}, \Lambda_i + \frac{1}{P_k + 0.05}) \\ \qquad \mathcal{N}(Y_{t+1} \mid \mu_i, \Sigma_i), & \text{if } d = 1. \\ \sum_{i=1}^{n} \sum_{k=1}^{K} w_k * \phi_{ki,2} \mathcal{N}^{-1}(X_{t+1} \mid \eta_i + \frac{m_k}{P_k + 0.05}, \Lambda_i + \frac{1}{P_k + 0.05}) \\ \qquad \mathcal{N}(Y_{t+1} \mid \mu_i, \Sigma_i), & \text{if } d = 2. \end{cases} \tag{99}$$

F.4.1 Implicit Reward

$$\hat{I}(X_{t+1}; Y_{t+1}) = H_{\hat{p}}(X_{t+1}) - H_{\hat{p}}(X_{t+1} \mid Y_{t+1}) \tag{100}$$

MI estimation: Given the Eq.(99) and normalize the weights, we could moment-match it to a single Gaussian distribution. Then we could compute $\hat{I}(X_{t+1}; Y_{t+1})$ analytically.

Posterior update: Given $Y_{t+1} = y_{t+1}$,

$$\hat{p}(X_{t+1} \mid y_{t+1}) \propto \begin{cases} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \phi_{ki,0} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i) \\ \qquad \mathcal{N}(X_{t+1} \mid \frac{\eta_i + \frac{m_k - 2}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}, \frac{1}{\Lambda_i + \frac{1}{P_k + 0.05}}), & \text{if } d = 0. \\ \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \phi_{ki,1} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i) \\ \qquad \mathcal{N}(X_{t+1} \mid \frac{\eta_i + \frac{m_k + 2}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}, \frac{1}{\Lambda_i + \frac{1}{P_k + 0.05}}), & \text{if } d = 1. \\ \sum_{k=1}^{K} w_k \sum_{i=1}^{n} \phi_{ki,2} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i) \\ \qquad \mathcal{N}(X_{t+1} \mid \frac{\eta_i + \frac{m_k}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}, \frac{1}{\Lambda_i + \frac{1}{P_k + 0.05}}), & \text{if } d = 2. \end{cases} \tag{101}$$

Fit a Gaussian $\mathcal{N}(X_{t+1}) \approx \sum_{i=1}^{n} \phi_{ki,d} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i)\mathcal{N}(X_{t+1} \mid \cdot)$ by moment-matching. The weight $w_k^{(new)} = \frac{w_k * \sum_{i=1}^{n} \phi_{ki,d} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i)}{\sum_{k=1}^{K} w_k * \sum_{i=1}^{n} \phi_{ki,d} \mathcal{N}(Y_{t+1} = y_{t+1} \mid \mu_i, \Sigma_i)}$

F.4.2 Explicit Reward

The explicit reward function is a function of the state s given a decision d, i.e.,

$$r(X) = \begin{cases} -2\mathcal{N}(X \mid -21, 1) - 2\mathcal{N}(X \mid -19, 1) - 2\mathcal{N}(-X \mid 17, 1), & \text{if } d = 0. \\ -2\mathcal{N}(X \mid 21, 1) - 2\mathcal{N}(X \mid 19, 1) - 2\mathcal{N}(X \mid 17, 1), & \text{if } d = 1. \\ -10\mathcal{N}(X \mid -25, 250) + 2\mathcal{N}(X \mid 3, 3) - 10\mathcal{N}(X \mid 25, 250), & \text{if } d = 2. \end{cases} \tag{102}$$

Given a decision d, denote the reward as $r_d(X) = \sum_i w_i^{(d)} \phi(X \mid \mu_i^{(d)}, \Sigma_i^{(d)})$. To note, the reward function is a linear combination of Gaussians w.r.t the state instead of a normalized Gaussian mixture model. In the POMDP environment, we are not able to evaluate the exact reward function but compute the expected reward w.r.t the augmented distribution, Let $\hat{\mu}_{t+1}^{(0)} = \frac{\eta_i + \frac{m_k - 2}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}$,$\hat{\mu}_{t+1}^{(1)} = \frac{\eta_i + \frac{m_k + 2}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}$,$\hat{\mu}_{t+1}^{(2)} = \frac{\eta_i + \frac{m_k}{P_k + 0.05}}{\Lambda_i + \frac{1}{P_k + 0.05}}$, and $\hat{var}_{t+1} = \Lambda_i + \frac{1}{P_k + 0.05}$

$$< r_d, b > = \int \int \sum_i w_i^{(d)} \phi(X_{t+1} \mid \mu_i^{(d)}, \Sigma_i^{(d)}) \hat{p}_d(x_{t+1}, y_{t+1}) \mathrm{d}x_{t+1} \mathrm{d}y_{t+1} \tag{103}$$