Efficient Variational Sequential Information Control

Jianwei Shen SJWJAMES@ARIZONA.EDU
Department of Computer Science
The University of Arizona
Tucson, AZ 85721, USA

Jason Pacheco PACHECOJ@CS.ARIZONA.EDU
Department of Computer Science
The University of Arizona
Tucson, AZ 85721, USA

Abstract

We develop a family of fast variational methods for sequential control in dynamical 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.

1. Introduction

Real world optimal experimental design (Blackwell, 1950; Bernardo, 1979) requires fast and flexible uncertainty quantification in complex environments. Pioneering work by Lindley (Lindley, 1956) suggests a Bayesian approach that maximizes mutual information (MI) (Cover and Thomas, 2006; MacKay et al., 2003), which lacks a closed-form in many real-world settings (Paninski, 2003). Existing work approximates MI using variational methods (Poole et al., 2019), Monte Carlo (Drovandi et al., 2013, 2014; Solonen et al., 2012), discretizations (Kim et al., 2014), and explore these approximations in a BOED setting (Huan and Marzouk, 2016; Kleinegesse and Gutmann, 2019).

This work extends variational BOED (Pacheco and Fisher, 2019; Foster et al., 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). Our setting subsumes a range of applications, e.g., active simultaneous localization and mapping (Active SLAM) (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). MI is also an effective intrinsic reward for RL tasks when extrinsic rewards are sparse or lacking (Fischer and Tas, 2020; Mohamed and Jimenez Rezende, 2015).

We address the computational aspects of control in this work by developing variational techniques for approximate inference and decision-making. We numerically evaluate our methods in different experiments. In all cases we observe multiple orders of magnitude speedup over sequential Monte Carlo (SMC) and accuracy comparable to numerical evaluation.

2. Information control

We formulate optimal information control as an instance of stochastic control with MI rewards. We consider an optimal control problem for the dynamical system with latent variables $X_0^T = X_0, \ldots, X_T$, observations Y_1^T, and joint PDF: $p(X_0^T, Y_1^T \mid d_1^T) = p(X_0) \prod_{t=1}^T p(X_t \mid X_{t-1}, d_t) p(Y_t \mid X_t)$. 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 MI over the sequence:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\ I(X_1^T; Y_1^T \mid \pi). \tag{1}$$

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. Furthermore, the global MI control objective in Eqn. (1) decomposes as a sum of conditional MI terms:

$$\max_{\pi}\ I(X_1; Y_1 \mid \pi) + \sum_{t=2}^{T} I(X_t; Y_t \mid Y_1^{t-1}, \pi) \tag{2}$$

A derivation of the above is provided in the Appendix.

Greedy Information Control A key property of Eqn. (2) is that each term depends on only a single latent state X_t. This suggests a simple greedy approximation at each time t:

$$d_t^* = \underset{d}{\operatorname{argmax}}\, H(X_t \mid \mathcal{H}_{t-1}, d) - H(X_t \mid Y_t, \mathcal{H}_{t-1}, d) = \underset{d}{\operatorname{argmax}}\, \mathbb{E}\left[\log \frac{p(X_t \mid Y_t, \mathcal{H}_{t-1}, d)}{p(X_t \mid \mathcal{H}_{t-1}, d)}\right]. \tag{3}$$

We denote the history data until time t as $\mathcal{H}_{t-1} = \{y_1^{t-1}, d_1^{t-1}\}$, where y_1^{t-1} are realized measurements and d_1^{t-1} are decisions taken. This dependence on realized observations induces a closed-loop sequential decision-making process. This greedy approach is not optimal in general, but it yields efficient high-quality approximate decisions in complex environments.

Calculating MI reward The instantaneous MI, or the greedy objective (Eqn. (3)) is not directly observed and cannot be computed in most settings (Mafi et al., 2011; Still and Precup, 2012; Mazzaglia et al., 2022). A simple approach to approximating MI in this setting is via a nested Monte Carlo (NMC) estimator. Given joint posterior samples $\{(x_t^i, y_t^i)\}_{i=1}^N \sim p(X_t, Y_t \mid \mathcal{H}_{t-1}, d)$ we have the NMC estimate: $\hat{I}_N = \frac{1}{N} \sum_{i=1}^N \log \frac{p(y_t^i \mid x_t^i)}{\frac{1}{N-1} \sum_{j \neq i} p(y_t^i \mid x_t^j)}$. NMC estimators of MI are consistent, asymptotically unbiased, and admit a central limit theorem. However, they require posterior samples and exhibit significant finite sample bias that decays slowly (Zheng et al., 2018; Rainforth et al., 2018) making them impractical in many settings.

3. Variational MI control

We detail our variational approach to greedy information control in a general context. We start by motivating our approach with existing work in sequential Bayesian optimal experimental design (BOED) (Pacheco and Fisher, 2019; Foster et al., 2018, 2019). We then describe the difficulties of time-varying latent variables in the dynamic control problem and extensions of BOED that require alternative variational approximations for the MI objective. After that, we provide details of our approach including the use of *assumed density filtering (ADF)* and *expectation propagation (EP)* inference. We show how these mechanisms yield computationally efficient variational MI approximations for control. We drop the history data $\mathcal{H}_{t-1}$ unless necessary to reduce the notation for the rest of the paper.

3.1 Variational BOED to greedy information control

The (greedy) sequential BOED objective optimizes instantaneous MI as $d_t^* = \text{argmax}_{d_t}\ H(X) - H(X \mid Y_t, d_t)$. Agakov and Barber (2004) lower bound MI by Gibbs' inequality while other work explores upper bounds (Poole et al., 2019; Foster et al., 2019). We build on these ideas in the sequel to derive variational control in the dynamic setting. Variations of BOED exist where X is time-varying (e.g. X_t) but the decision variable modulates only the observation model (Williams et al., 2007; Shamaiah et al., 2010).

Unlike the static BOED setting, the control model of Sec. 2 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. (3) involve the control variate d_t, and neither can be computed in closed-form. We replace both terms with cross-entropies over variational distributions,

$$I(X_t; Y_t \mid d_t) \approx H_{p_{d_t}}(q(X_t \mid d_t)) - H_{p_{d_t}}(q(X_t \mid Y_t, d_t)). \tag{4}$$

The approximation in Eqn. (4) is neither an upper- nor a lower-bound on MI, but is instead an approximation. This approximation was previously explored in the context of implicit likelihood models (Foster et al., 2019). Despite the simple form of the approximation in Eqn. (4), optimizing and evaluating it remains challenging, which we address next.

3.2 Estimating the variational MI approximation

The entropy terms in the variational approximation of instantaneous MI (Eqn. (4)) require expectations w.r.t. the joint posterior over state and measurement $p(X_t, Y_t \mid d_t)$. This distribution is not available in practice so we perform approximate variational inference via *assumed density filtering (ADF)* (Murphy, 2012). Fig. 1 provides a depiction of the stages of ADF inference and MI approximation in our method.

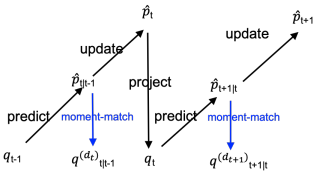


Figure 1: ADF and MI evaluation.

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 Eqn. (5). This augmented distribution is a local approximation to the predictive distribution and is not an exponential

family in general so its entropy cannot be calculated easily.

$$\hat{p}_{t|t-1}(X_t, Y_t \mid d_t) = p(Y_t \mid X_t) \int q_{t-1}(X_{t-1}) p(X_t \mid X_{t-1}, d_t) \, \mathrm{d}X_{t-1}. \tag{5}$$

Moment-matching Step Given the augmented distribution in Eqn. (5) the greedy MI surrogate objective becomes,

$$I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid d_t) \approx H_{\hat{p}_{t|t-1}}(q_m(X_t)) - H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) \equiv I_{\hat{p}_{t|t-1}}(q). \tag{6}$$

Cross entropy is an expectation w.r.t. the augmented distribution $\hat{p}_t$ instead of the true filter as in Eqn. (3). Finding the optimal variational distribution for Eqn. (6) is a non-convex optimization. We take the approach of (Foster et al., 2019) and minimize an upper bound on the absolute error,

$$|I_{\hat{p}_{t|t-1}} - I_{\hat{p}_{t|t-1}}(q)| \leq \min_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) + \min_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) + C \tag{7}$$

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.

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.* (7).

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 Eqn. (6) is closed-form at the moment-matching solution.

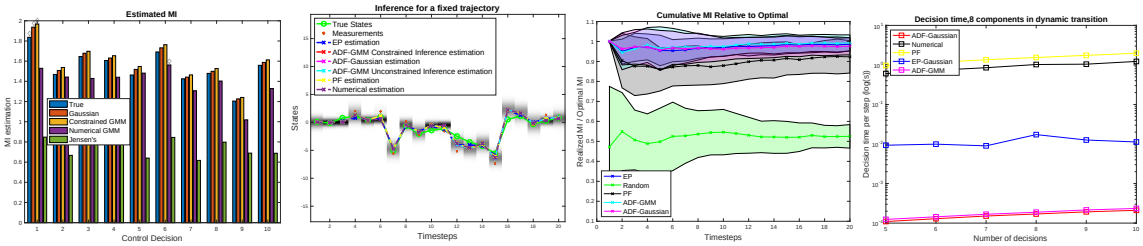
Update Step. After the control d_t with maximum (approximate) MI, empirically evaluated via Eqn. (6), is selected and executed, a realized measurement is obtained $y_t \sim p(Y_t \mid X_t)$ from the environment. Then, an ADF update is performed to yield an exponential family approximation $q_t \approx \hat{p}_t(X_t \mid \mathcal{H}_{t+1})$ via KL-projection (e.g. moment-matching in the exponential family). Moreover, we extend the ADF-driven variational approach to the Expectation propagation (EP) inference for more accurate inference results. We demonstrate the accuracy of inference for both the ADF and EP in Fig. 2b, under a fixed decision-state-measurement trajectory to compare with the baseline and the ground truth.

4. Information control for GMM-Gaussian dynamical systems

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

$$p(X_0^T, Y_1^T \mid d_1^T) = \mathcal{N}(X_0 \mid m_0, \Sigma_0) \prod_{t=1}^{T} \mathcal{N}(Y_t \mid FX_t, R) \sum_{k=1}^{K} w_{k,d_t} \mathcal{N}(X_t \mid A_{k,d_t} X_{t-1}, Q_{k,d_t}). \tag{8}$$

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, which is widely used for model-based learning of dynamical systems (Khansari-Zadeh



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

Figure 2: (a) One-step decision-making experiment. MI estimations of the $K^2 = 16$-component GMM by five methods, with diamond character marking the decision each method will take respectively. It also validates our theorem on the upper bound **yellow** $\geq$ **orange** $\geq$ **blue**. (b) Posterior inference of the same decision trajectory for different methods, grayscale bars representing the filter distribution given by the numerical approximation method (approximate the true filter distribution) (c) Ratio of cumulative MI against the optimal MI estimated by numerical approximation (d) Runtime v.s. the number of decisions, with each decision bounded to an 8-component GMM dynamic transition.

and Billard, 2011; Hersch et al., 2008). In Sec. 4.1, we include an application of the general solution in Sec. 3.2 to this model with Gaussian approximation, as well as an upper bound on the MI. Besides, we propose the development of a constrained Gaussian mixture variational family in Sec. 4.2, which allows the analytic calculation of MI under the variational projection. We observe that our methods indicate orders of magnitude speedup in decision-making when in an online sequential decision-making process(c.f. Fig. 2d).

4.1 Gaussian MI approximation

We apply ADF and EP as discussed in Sec. 3.2 with standard Gaussian variational approximations. We refer to these baseline Gaussian methods as simply "ADF-Gaussian" and "EP" in experimental results of Sec. 5. Following the **prediction step** in Sec. 3.2 conducted under this model, we can upper bound the MI w.r.t the augmented distribution $\hat{p}_{t|t-1}(X_t, Y_t \mid d_t)$ by its Gaussian approximation, $q^{(d_t)}_{t|t-1}(X_t, Y_t)$, i.e., $I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid d_t) \leq I_{q^{(d_t)}_{t|t-1}}(X_t; Y_t \mid d_t)$.

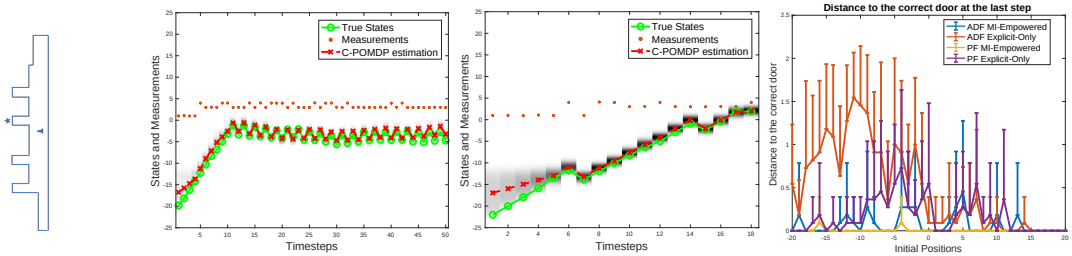
We officially state and prove this proposition in the Appendix. In fact, it holds independent of the state dynamical transitions, provided the measurement model is a linear Gaussian distribution. Note that the bound is w.r.t. the local approximation under the augmented distribution as opposed to the true filter. To better adapt to the true GMM filter distribution we develop a specialized approach based on GMM projections outlined next.

4.2 Constrained GMM MI approximation

At each time $t-1$, we maintain an ensemble of K Gaussians, $q_{t-1}(X_{t-1} \mid S_{t-1} = k) = \mathcal{N}(X_{t-1} \mid \hat{\mu}_{k,t-1}, \hat{\Sigma}_{k,t-1})$ with $q_{t-1}(S_{t-1} = k) = \pi_{k,t-1}$. The posterior distribution is $q_{t-1}(X_{t-1}) = \sum_{i=1}^{K} \hat{\pi}_{i,t-1} \mathcal{N}(X_{t-1} \mid \hat{\mu}_{i,t-1}, \hat{\Sigma}_{i,t-1})$.

Prediction step. This yields a K^2-component GMM augmented distribution for each control variate, denoted as $\hat{p}_{t|t-1}(X_t, Y_t, S_t = k \mid d_t)$.

Constrained GMM moment-matching step. We constrain the variational approximation to factorize as: $q^{(d_t)}_{t|t-1}(X_t, Y_t, S_t = k) = \omega_k \mathcal{N}(Y_t|\eta, P)\mathcal{N}(X_t|F_k Y_t, M_k)$. The Gaussian



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

Figure 3: (a) Environment with the target (star) and agent. (b) The exact-reward-only mode can get stuck in an oscillation. (c) MI-empowered reward and ADF inference find the target. (d) 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 in reaching the correct door as the near-optimal PF approximation with an MI-empowered reward. And it is more stable than the explicit-only methods.

marginal on Y is invariant to mixture component allowing marginal moments η and P to be computed directly from the target distribution. GMM MI cannot be easily calculated in general but our constrained GMM ensures that the MI has a simple form, which can be verified by direct calculation. We minimize KL by moment-matching and compute the corresponding MI of the moment-matched constrained GMM distribution. Note that a fixed-component Gaussian ensemble is in the exponential family thus ADF/EP updates are easily computed (Heskes and Zoeter, 2002; Pacheco and Sudderth, 2012). We therefore have a surrogate MI objective,

$$I_{q_{t|t-1}^{(d_t)}}(\{X_t, S_t\}; Y_t) = \sum_{k=1}^{K} \omega_k \left[H_{q_{t|t-1}^{(d_t)}}(X_t \mid S_t = k) - H_{q_{t|t-1}^{(d_t)}}(X_t \mid Y_t, S_t = k) \right]. \tag{9}$$

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. Moreover, we have a useful upper bound, $I_{q_{t|t-1}^{(d_t)}}(\{X_t, S_t\}; Y_t) \geq I_{q_{t|t-1}^{(d_t)}}(X_t; Y_t) \geq I_{\hat{p}_{t|t-1}}(X_t; Y_t \mid d_t)$. Our surrogate does not bound the desired true posterior MI $I_{p_{t|t-1}}$, but performs well empirically (c.f. Fig. 2a). We observe that our methods yield comparatively high-quality decision selections in a sequential decision-making process, which leads to near-optimal MI evaluation, against PF (Particle Filtering) baseline, presented as Fig. 2c.

5. MI-empowered one-shot continuous POMDP

We adopt the basic environmental setup as Porta et al. (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. 3a. We defer the detailed settings to the appendix because of the space limitation, but we want to note that, keeping track of the *belief state*, i.e., the posterior distribution $b_t = p(X_t \mid \mathcal{H}_{t+1})$ is infeasible since it is a GMM and the number of components grows exponentially. We applied the method introduced in the previous sections to efficiently address the intractable inference problem and the decision-making problem to achieve the target. We illustrate part of the results from Fig. 3b to Fig. 3d.

References

Felix Agakov and David Barber. The IM algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

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.

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

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.

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

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

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

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.

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

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.

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.

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.

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

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.

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.

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.

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.

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

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.

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.

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

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.

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

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

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.

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

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.

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

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.

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

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.

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

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.

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.

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

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

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.

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.

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.

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.

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

Jason L. Williams, John W. Fisher, and Alan S. Willsky. Approximate dynamic programming for communication-constrained sensor network management. *IEEE Transactions on Signal Processing*, 55(8):4300–4311, 2007. doi: 10.1109/TSP.2007.896099.

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.

Appendix A. Global MI control objective

For convenience, we restate the MI control objective from the main text (Eqn. (1)) here:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\ I(X_1^T; Y_1^T \mid \pi). \tag{10}$$

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. (2). 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:

$$I(A; \{B, C\}) = I(A; B) + I(A; C|B), \tag{11}$$

By the MI chain rule on variables Y_1^T the MI control objective in Eqn. (10) decomposes additively as:

$$\begin{aligned} I(X_1^T; Y_1^T) &= I(X_1^T; Y_1) + I(X_1^T; Y_2^T|Y_1) \\ &= I(X_1^T; Y_1) + I(X_1^T; Y_2|Y_1) + I(X_1^T; Y_3^T|Y_1) \\ &\dots \\ &= I(X_1^T; Y_1) + \sum_{t=2}^{T} I(X_1^T; Y_t|Y_1^{t-1}). \end{aligned} \tag{12}$$

The chain rule is further applied on variables X_1^T for each term in Eqn. (12). Taking the first term as an example we have,

$$I(X_1^T; Y_1) = I(X_1; Y_1) + \sum_{t=2}^{T} I(X_t; Y_1|X_1^{t-1}) = I(X_1; Y_1). \tag{13}$$

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. (12) simplifies as,

$$I(X_1^T; Y_t \mid Y_1^{t-1}) = I(X_t; Y_t \mid Y_1^{t-1}) + I(\{X_i\}_{i\in\{1,\dots,T\}\setminus t}; Y_t \mid Y_1^{t-1}, X_t) = I(X_t; Y_t \mid Y_1^{t-1}). \tag{14}$$

Combining these steps we have the decomposed global MI objective,

$$I(X_1^T; Y_1^T) = I(X_1; Y_1) + \sum_{t=2}^{T} I(X_t; Y_t|Y_1^{t-1}). \tag{15}$$

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. (15). 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.

Appendix B. Moment-matching in Gaussian case

We propose the Theorem 3.1 and claim that Eqn. (6) 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

For the Gaussian case, we claim that $q_{t|t-1}^{(d_t)}(X_t, Y_t) = \mathcal{N}(m, \Sigma)$ by moment-matching $\hat{p}_{t|t-1}(X_t, Y_t)$ yields optimal Gaussian $q_m^* = \int q_{t|t-1}^{(d_t)}(X_t, y_t)\mathrm{d}y_t$ and $q_c^* = \frac{q_{t|t-1}^{(d_t)}(X_t, Y_t)}{\int q_{t|t-1}^{(d_t)}(x_t, Y_t)\mathrm{d}x_t}$.

Proof Recall the upper bound on MI error as Eqn. (7),

$$|I_{\hat{p}_{t|t-1}} - I_{\hat{p}_{t|t-1}}(q)| \leq \min_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) + \min_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) + C. \tag{16}$$

For Gaussian marginal and conditional,

$$q_m(X_t) = \mathcal{N}(\mu, Q) \quad \text{and} \quad q_c(X_t \mid Y_t) = \mathcal{N}(AY_t + b, \Gamma), \tag{17}$$

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

$$q_m^* = \operatorname*{argmin}_{q_m} H_{\hat{p}_{t|t-1}}(q_m(X_t)) - \underbrace{H_{\hat{p}_{t|t-1}}(X_t)}_{\text{constant}} = \operatorname*{argmin}_{q_m} \mathrm{KL}(\hat{p}_{t|t-1} \,\|\, q_m), \tag{18}$$

so

$$\mu^* = E_{\hat{p}_{t|t-1}}[X_t] \quad \text{and} \quad Q^* = \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t). \tag{19}$$

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:

$$\begin{aligned} q_c^* &= \operatorname*{argmin}_{q_c} H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)) \\ \alpha(A, b, \Gamma) &\equiv \min_{A,b,\Gamma} E\left[-\log \mathcal{N}(X_t \mid AY_t + b, \Gamma)\right] \\ &= \min_{A,b,\Gamma} \frac{1}{2}\log \mid \Gamma \mid + E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right]. \end{aligned} \tag{20}$$

First, we solve for b,

$$\begin{aligned} \nabla_b \alpha &= \nabla_b E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \\ &= E\left[\Gamma^{-1}(X_t - b - AY_t)(-\nabla_b b)\right] \\ &= -E\left[\Gamma^{-1}(X_t - b - AY_t)\right] = 0 \\ b^* &= E\left[(X_t - AY_t)\right]. \end{aligned} \tag{21}$$

Second, we solve for A,

$$\begin{aligned}
\nabla_A \alpha &= \nabla_A E\left[\frac{1}{2} tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] = 0 \\
0 &= E\left[X_t Y_t^T - b Y_t^T - A Y_t Y_t^T\right] \\
&\textbf{Backsubstitute} \quad b^* \\
0 &= E\left[X_t Y_t^T - E\left[(X_t - AY_t)\right] Y_t^T - A Y_t Y_t^T\right] \\
&= E\left[X_t Y_t^T\right] - E\left[X_t\right] E\left[Y_t^T\right] + AE\left[Y_t\right] E\left[Y_t^T\right] - AE\left[Y_t Y_t^T\right] \\
&= \mathrm{Cov}(X_t, Y_t) - A\mathrm{Cov}(Y_t, Y_t) \\
A^* &= \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}
\end{aligned} \tag{22}$$

Back substitute into b^*,

$$b^* = E\left[X_t - AY_t\right] = E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1} E\left[Y_t\right]. \tag{23}$$

The full conditional mean is then,

$$\begin{aligned}
A^* Y_t + b^* &= \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1} Y_t + E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1} E\left[Y_t\right] \\
&= E\left[X_t\right] + \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right])
\end{aligned} \tag{24}$$

Solve for conditional covariance Γ,

$$\begin{aligned}
\nabla_\Gamma \alpha &= \nabla_\Gamma \frac{1}{2} \log \mid \Gamma \mid + \frac{1}{2} E\left[tr(\Gamma^{-1}(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] \\
&= \Gamma - E\left[(X_t - b - AY_t)(X_t - b - AY_t)^T)\right] = 0 \\
\Gamma^* &= E\left[(X_t - b - AY_t)(X_t - b - AY_t)^T)\right]
\end{aligned} \tag{25}$$

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

$$\begin{aligned}
\Gamma^* = E[&(X_t - E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right]) \\
&(X_t - E\left[X_t\right] - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}(Y_t - E\left[Y_t\right])^T] \\
&= \mathrm{Cov}(X_t, X_t) - \mathrm{Cov}(X_t, Y_t)\mathrm{Cov}(Y_t, Y_t)^{-1}\mathrm{Cov}(X_t, Y_t)^T.
\end{aligned} \tag{26}$$

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

$$\begin{aligned}
m &= \begin{bmatrix} m_X \\ m_Y \end{bmatrix} = \begin{bmatrix} E_{\hat{p}_{t|t-1}}[X_t] \\ E_{\hat{p}_{t|t-1}}[Y_t] \end{bmatrix} \\
\Sigma &= \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, X_t) & \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, Y_t) \\ \mathrm{Cov}_{\hat{p}_{t|t-1}}(X_t, Y_t)^T & \mathrm{Cov}_{\hat{p}_{t|t-1}}(Y_t, Y_t) \end{bmatrix}
\end{aligned} \tag{27}$$

Thus,

$$q_m = \mathcal{N}(m_X, \Sigma_X) = \mathcal{N}(\mu^*, Q^*) = \underset{q_m}{\mathrm{argmin}}\, H_{\hat{p}_{t|t-1}}(q_m(X_t)). \tag{28}$$

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

$$q_c = \mathcal{N}(X_t \mid \underbrace{m_X + \Sigma_{XY}\Sigma_Y^{-1}(Y_t - m_Y)}_{A^* Y_t + b^*}, \underbrace{\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T}_{\Gamma^*}) = \underset{q_c}{\mathrm{argmin}}\, H_{\hat{p}_{t|t-1}}(q_c). \tag{29}$$

Therefore, moment-matching the augmented distribution to a joint Gaussian distribution yields optimal Gaussian marginal and conditional approximations that minimize an upper bound on the error. ■

B.2 Closed-form solution for MI

Having computed the moment matching distribution q we must compute the MI approximation as,

$$I_{\hat{p}_{t|t-1}}(q) \equiv H_{\hat{p}_{t|t-1}}(q_m(X_t)) - H_{\hat{p}_{t|t-1}}(q_c(X_t \mid Y_t)). \tag{30}$$

It is not immediately obvious that these cross-entropies can be calculated in closed-form. However, it can be showed that for a moment-match Gaussian distribution, each entropy term can be easily calculated. In fact the cross-entropy terms equate to Gaussian entropy.

Theorem B.1. *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

$$I_{\hat{p}_{t|t-1}}(q) = H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) - H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t)). \tag{31}$$

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 the dimension of Σ_X is k,

$$\begin{aligned} H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) &= -E_{\hat{p}_{t|t-1}}[\log \mathcal{N}(X_t \mid m_X, \Sigma_X)] \\ &= -E_{\hat{p}_{t|t-1}}[-\frac{k}{2}\log 2\pi - \frac{1}{2}\log \mid \Sigma_X \mid -\frac{1}{2}(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] \\ &= \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X \mid + \frac{1}{2} E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] \end{aligned} \tag{32}$$

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

$$E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] = E_{\hat{p}_{t|t-1}}[(L^{-1}X_t - L^{-1}m_X)^T (L^{-1}X_t - L^{-1}m_X)] \tag{33}$$

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,

$$C = \mathrm{Cov}(L^{-1}X_t, L^{-1}X_t) = L^{-1} \underbrace{\mathrm{Cov}(X_t, X_t)}_{\Sigma_X} L^{-T} = I. \tag{34}$$

Therefore,

$$E_{\hat{p}_{t|t-1}}[(X_t - m_X)^T \Sigma_X^{-1}(X_t - m_X)] = tr(C) = k. \tag{35}$$

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

$$H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t)) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X \mid + \frac{k}{2}. \tag{36}$$

A similar result applies to H, but replace Σ_X with $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T$

$$H_{\hat{p}_{t|t-1}}(q_{t|t-1}(X_t \mid Y_t)) = \frac{k}{2}\log 2\pi + \frac{1}{2}\log \mid \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \mid + \frac{k}{2}. \tag{37}$$

$$I_{\hat{p}_{t|t-1}}(q) = \frac{1}{2}\log \mid \Sigma_X \mid - \frac{1}{2}\log \mid \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^T \mid . \tag{38}$$

■

Appendix 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 $x_0 \sim N(\mu_0, \Sigma_0)$
Output: A series of decisions $D_{1:T}$
Initialization: $\alpha_0(x_0) = N(\mu_0, \Sigma_0), \beta_{0:T-1} = 1$
Let $D[t]$ be the optimal decision at time t when the final step is t,$1 \leq t \leq T$
{// 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
  $d_t^* = \text{argmax}_{d_t \in \{1,\ldots,K\}} \; I_{\hat{p}_{t,d_t}}(q_{t,d_t})$ {// 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

Appendix D. Moment-matching calculation of constrained GMM

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

$$p_d(X_{t+1}|X_t) = \sum_{i=1}^N w_d(i) N(X_{t+1}|A_{d,i}X_i, C_{d,i}), \tag{39}$$

$$p(Y_{t+1}|HX_{t+1}, R), \tag{40}$$

for a given decision d

$$\begin{aligned}
\hat{p}(X_{t+1}, Y_{t+1}, S_{t+1} = j) &= \int \sum_{i=1}^K \pi_i N(X_t|\mu_i, \Sigma_i) w_j N(X_{t+1}|A_j X_t, C_j) p(Y_{t+1}|HX_{t+1}, R) d_{X_t} \\
&= \sum_{i=1}^K \pi_i w_j N(Y_{t+1}|HX_{t+1}, R) N(X_{t+1} \mid A_j\mu_i, C_j + A_j\Sigma_i A_j^T) \\
&= \sum_{i=1}^K \pi_i w_j N(Y_{t+1}|HX_{t+1}, R) N(X_{t+1} \mid m_{ij}, P_{ij})
\end{aligned} \tag{41}$$

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

$$q(X_{t+1}, Y_{t+1}, S_{t+1} = j) = N(Y_{t+1}|\eta, P) w_j N(X_{t+1}|F_j Y_{t+1}, M_j). \tag{42}$$

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

$$\begin{aligned}
\hat{p}(Y_{t+1}) &= \sum_{i=1}^K \pi_i \sum_{j=1}^N w_j \int N(Y_{t+1} \mid Hx_{t+1}, R) N(x_{t+1}|m_{ij}, P_{ij}) dx_{t+1} \\
&= \sum_{i=1}^K \pi_i \sum_{j=1}^N w_j N(Y_{t+1} \mid Hm_{ij}, R + HP_{ij}H^T)
\end{aligned} \tag{43}$$

Let

$$\begin{aligned}
M &= \sum_{i=1}^K \pi_i \sum_{j=1}^N w_j (Hm_{ij}), \\
V &= \sum_{i=1}^K \pi_i \sum_{j=1}^N w_j [R + HP_{ij}H^T + (Hm_{ij})(Hm_{ij})^T] - MM^T
\end{aligned} \tag{44}$$

$$\begin{aligned}
P &= \mathrm{Cov}_{\hat{p}}(Y) = V \\
\eta &= M
\end{aligned} \tag{45}$$

To compute F_j and M_j, we consider

$$\begin{aligned}
G(F_j, M_j) &\equiv \min_{F_j, M_j} E\left[-\log \mathcal{N}(X_t \mid F_j Y_t, M_j)\right] \\
&= \min_{F_j, M_j} \frac{1}{2} \log \mid M_j \mid + E\left[\frac{1}{2} tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right].
\end{aligned} \tag{46}$$

Solve for F_j,

$$\begin{aligned}
\nabla_{F_j} G &= \nabla_{F_j} E\left[\frac{1}{2} tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\
&= E\left[M_j^{-1}(X_t - F_j Y_t)(-\nabla_{F_j} F_j Y_t)\right] \\
&= -E\left[M_j^{-1}(X_t - F_j Y_t) Y_t^T\right] = 0 \\
0 &= E\left[X_t Y_t^T - F_j Y_t Y_t^T\right] \\
F_j^* &= E\left[X_t Y_t^T\right] E\left[Y_t Y_t^T\right]^{-1} \\
&= \{\mathrm{Cov}(X_t, Y_t) + E[X_t]E[Y_t^T]\}\{\mathrm{Cov}(Y_t, Y_t) + E[Y_t]E[Y_t^T]\}^{-1}
\end{aligned} \tag{47}$$

Solve for M_j,

$$\begin{aligned}
\nabla_{M_j} G &= \nabla_{M_j} \frac{1}{2} \log \mid M_j \mid + \frac{1}{2} E\left[tr(M_j^{-1}(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\
&= M_j - E\left[(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] = 0 \\
M_j^* &= E\left[(X_t - F_j Y_t)(X_t - F_j Y_t)^T)\right] \\
&= \mathrm{Cov}(X_t, X_t) + E[X_t]E[X_t^T] - \{\mathrm{Cov}(X_t, Y_t) + E[X_t]E[Y_t^T]\} \\
&\{\mathrm{Cov}(Y_t, Y_t) + E[Y_t]E[Y_t^T]\}^{-1}\{\mathrm{Cov}(Y_t, X_t) + E[Y_t]E[X_t^T]\}
\end{aligned} \tag{48}$$

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

$$\hat{p}(X_{t+1}, Y_{t+1}, S_{t+1} = j) = \Sigma_{i=1}^K \pi_i w_j N(Y_{t+1}|HX_{t+1}, R) N(X_{t+1}|m_{ij}, P_{ij}) \tag{49}$$

$$\begin{aligned}
\hat{p}(X_{t+1}, Y_{t+1}|S_{t+1} = j) &= \Sigma_{i=1}^K \pi_i N(Y_{t+1}|HX_{t+1}, R) N(X_{t+1}|m_{ij}, P_{ij}) \\
&= \Sigma_{i=1}^K \pi_i N\left(\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} \middle| \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix}, \begin{bmatrix} P_{ij} & P_{ij}H^T \\ HP_{ij} & R + HP_{ij}H^T \end{bmatrix}\right)
\end{aligned} \tag{50}$$

Let

$$\begin{aligned}
\widetilde{\mu}_{t+1} &= \Sigma_{i=1}^K \pi_i \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix}, \\
\widetilde{V}_{t+1} &= \Sigma_{i=1}^K \pi_i \left\{ \begin{bmatrix} P_{ij} & P_{ij}H^T \\ HP_{ij} & R + HP_{ij}H^T \end{bmatrix} + \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix} \begin{bmatrix} m_{ij} \\ Hm_{ij} \end{bmatrix}^T \right\} \\
&-\widetilde{\mu}_{t+1}\widetilde{\mu}_{t+1}^T
\end{aligned} \tag{51}$$

Let

$$m_j = \Sigma_{i=1}^K \pi_i m_{ij}, V_j = \Sigma_{i=1}^K \pi_i [P_{ij} + m_{ij}m_{ij}^T] - m_j m_j^T, \tag{52}$$

then

$$\mathrm{Cov}_{\hat{p}}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} V_j & V_j H^T \\ HV_j & R + HV_j H^T \end{bmatrix}, \tag{53}$$

$$\mathrm{Cov}_{\hat{p}}^{-1}(X_{t+1}, Y_{t+1}|S_{t+1} = j) = \begin{bmatrix} V_j^{-1} + H^T R^{-1} H & -H^T R^{-1} \\ -R^{-1}H & R^{-1} \end{bmatrix}, \tag{54}$$

$$\mathrm{Cov}_{\hat{p}}(X_{t+1}|S_{t+1}=j)=V_j. \tag{55}$$

Since

$$\mathrm{Cov}_{q}(X_{t+1},Y_{t+1}|S_{t+1}=j)=\begin{bmatrix} M_j+F_jPF_j^T & F_jP \\ PF_j^T & P \end{bmatrix} \tag{56}$$

and

$$\mathrm{Cov}_{q^{-1}}(X_{t+1},Y_{t+1}|S_{t+1}=j)=\begin{bmatrix} M_j^{-1} & -M_j^{-1}F_j \\ -F_j^TM_j^{-1} & P^{-1}+F_j^TM_j^{-1}F_j \end{bmatrix}, \tag{57}$$

$$\mathrm{Cov}_{q}(X_{t+1}|S_{t+1}=j)=M_j+F_jPF_j^T=V_j, \tag{58}$$

and

$$\mathrm{Cov}_{q}(X_{t+1}|Y_{t+1},S_{t+1}=j)=M_j. \tag{59}$$

Appendix E. MI-empowered one-shot POMDP learning

Environment and Challenge. We adopt the basic environmental setup as Porta et al. Porta et al. (2006) but focus on a one-shot POMDP learning problem to underline the need for **efficiency** and **accuracy**. 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 corridor width. The robot's objective is to reach the correct door by navigating the corridor (taking actions to move left, right, and enter), as Fig. 3a. 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. Given the belief state b_t and reward function $r(a, X)$, the optimal policy is learned by

$$\pi^* = \underset{\pi}{\mathrm{argmax}}\, E_{p(Y_t|\mathcal{H}_t)}[E_{b_t}[r(\pi(x), X_t = x) + \text{future rewards}]]. \tag{60}$$

A straightforward and effective method is to apply a greedy reward. But we observe that, with $E_{p(Y_t|\mathcal{H}_t)}[E_{b_t}[r(a_t, X_t = x)]]$ (greedy expected explicit reward) only, the performance of the robot is not stable and it tends to get stuck at one place as Fig. 3b.

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

$$R(a_t) = E_{p(Y_t|\mathcal{H}_t)}[E_{b_t}[r(a_t, X_t = x)]] + \alpha I(X_t; Y_t \mid a_t, \mathcal{H}_t). \tag{61}$$

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. 61. As a comparison, we also implement PF with 3000 samples in this space to approximate the truth. As shown in Fig. 3c, with MI-empowered reward, the robot has the ability to address the oscillation problem in Fig. 3b. 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. 3d) in 11 runs and calculate the mean and + 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.