Value of Information and Reward Specification in Active Inference and POMDPs

Ran Wei
VERSES
ran.wei@verses.ai

Abstract

Active inference is an agent modeling framework with roots in Bayesian predictive coding and the free energy principle. In recent years, active inference has gained popularity in modeling sequential decision making, a problem set traditionally populated by reinforcement learning (RL). Instead of optimizing expected reward as in RL, active inference agents optimize expected free energy (EFE), which has an intuitive decomposition into a pragmatic and an epistemic component. This makes us wonder: *what's the EFE-optimizing agent's optimality gap compared with a reward-driven RL agent, which is well understood?* By casting EFE under a particular class of belief MDP and using analysis tools from RL theory, we show that EFE approximates the Bayes optimal RL policy via information value. We discuss the implications for objective specification of active inference agents.

1 Introduction

Active inference (Parr et al., 2022) is an agent modeling framework with roots in Bayesian predictive coding (Friston et al., 2010, 2012) and the free energy principle (Friston, 2010). In recent years, active inference has seen increased popularity in various fields including but not limited to cognitive and neural science, machine learning, and robotics (Smith et al., 2021; Mazzaglia et al., 2022; Lanillos et al., 2021). One common application of active inference across these fields is in modeling decision making behavior, often taking place in partially observable Markov decision processes (POMDP) where active information gathering is crucial to task performance. This offers active inference as complementary, a potential alternative to, or a possible generalization of optimal control and reinforcement learning (RL).

The central difference between active inference and RL is that instead of choosing actions that maximize expected reward or utility, active inference agents are mandated to minimize expected free energy (EFE; Da Costa et al., 2020), which has an intuitive decomposition as the addition of a pragmatic value term and an epistemic value term. Notably, the epistemic value term encourages the agent to explore and gather information about the environment, a behavior primitive that is crucial in challenging partially observable task environments. Indeed, experimental evaluations of active inference agents have shown that the epistemic value term in EFE contributes to structured exploratory behavior, resolving uncertainty before attempting to obtain reward, often leading to higher coverage of the state space and enhanced task performance (Millidge, 2020; Tschantz et al., 2020; Engström et al., 2024).

It appears, at a first glance, that RL and optimal control miss the epistemic value term. However, it is widely known that the Bayes optimal policy in POMDPs already trades off exploration and exploitation (Roy et al., 2005). This makes intuitive sense because resolving uncertainty often leads to more downstream rewards, essentially by "opening up" opportunities. Specifically, the Bayes optimal policy leverages the equivalence between POMDPs and a special class of MDPs defined on the reward and transition of beliefs called *belief MDPs* to characterize the expected value (i.e., cumulative

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

reward) following an action given the current belief, from which an optimal policy can be constructed as a mapping from beliefs to actions (Kaelbling et al., 1998). These policies, as demonstrated by Bayes adaptive RL and meta RL, also exhibit structured exploratory behavior (Zintgraf et al., 2019; Duan et al., 2016). It thus begs the question: *What is the relationship between the Bayes optimal RL policy and the active inference policy based on optimizing EFE?*

The main contribution of this paper is providing one answer to the above question:
EFE approximates the Bayes optimal RL policy via epistemic value.
We achieve this by first establishing the equivalence between the EFE objective and a different class of belief MDPs, which allows us to define EFE-optimal policies to form direct comparisons with RL policies. We then examine the source of epistemic behavior in POMDPs using a definition of the value of information for POMDPs based on Howard's information value theory (1966). In brief, the value of information is the difference in the expected values between the Bayes optimal policy and another "naive" policy which plans as if it would not be able to update beliefs based on observations in the future. When casting the latter policy also using belief MDPs, we observe that it uses the same belief transition dynamics as the EFE policy but it uses the same belief reward as the Bayes optimal policy. Our key result is a regret bound showing that the EFE objective closes the performance gap between the naive policy and the Bayes optimal policy by augmenting or shaping the reward function of the former with epistemic value. We discuss the implications of our results for specifying active inference agents in practice.

2 Preliminaries

In this section, we introduce partially observable Markov decision process, value of information, and active inference and establish their corresponding belief MDPs.

2.1 Partially observable Markov decision process

A discrete time infinite-horizon discounted partially observable MDP (POMDP; Kaelbling et al., 1998) is defined by a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, R, \mu, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{O}$ are respectively finite sets of states, actions, and observations, P is the transition dynamics consisting of a state transition probability distribution $P(s_{t+1}|s_t, a_t)$ and an observation emission distribution $P(o_t|s_t)$, $R(s_t, a_t)$ is a scalar reward function, $\mu(s_0)$ the initial state distribution, and $\gamma \in (0, 1)$ a discount factor. The goal of an agent is to find a policy $\pi(a_t|h_t)$ mapping the interaction history $h_t = (o_{0:t}, a_{0:t-1})$ to a distribution over actions which maximizes the expected cumulative discounted reward: $J_M(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where the expectation is taken w.r.t. the stochastic process induced by the dynamics of environment M and agent policy. The process of finding the optimal policy is sometimes referred to as reinforcement learning (RL; Sutton and Barto, 2018). Importantly, the agent cannot observe the underlying environment state in the process.

It is a well-known result that the Bayesian belief distribution $b_t = P(s_t|h_t)$ is a sufficient statistic for the interaction history in POMDPs (Kaelbling et al., 1998). The history dependent value functions can thus be written in terms of beliefs which are treated as random variables:

$$Q(b, a) = \sum_s b(s)R(s, a) + \gamma \sum_{o'} P(o'|b, a)V(b'(o', a, b)), \quad V(b) = \max_a Q(b, a)\,, \tag{1}$$

where $P(o'|b, a) = \sum_{s,s'} P(o'|s')P(s'|s, a)b(s)$ and $b'(o', a, b)$ denotes the belief update function from prior $b(s)$ to the posterior:

$$b'(o', a, b) := b'(s'|o', a, b) = \frac{P(o'|s') \sum_s P(s'|s, a)b(s)}{\sum_{s'} P(o'|s') \sum_s P(s'|s, a)b(s)}\,. \tag{2}$$

The (Bayes) optimal policy can then be derived from the above value functions as $\pi(a|h) = \delta(a - \arg\max_{\tilde{a}} Q(b, \tilde{a}))$ where δ denotes the dirac delta distribution.

The belief value functions in (1) imply a special class of (fully observable) MDPs known as *belief MDPs* (Kaelbling et al., 1998), where the reward and dynamics are defined on the belief state as:

$$R(b, a) = \sum_s b(s)R(s, a), \quad P(b'|b, a) = P(o'|b, a)\delta(b' - \tilde{b}'(o', a, b))\,. \tag{3}$$

The stochasticity in the belief dynamics is entirely due to the stochasticity of the next "counterfactual" observation; the belief updating process itself is deterministic.

In this work, we generalize the notion of belief MDP to refer to any MDP defined on the space of beliefs. However, not all belief MDPs could yield the optimal policies for some POMDPs.

2.2 Value of information

It is colloquially accepted that the Bayes optimal POMDP policy trades of exploration and exploitation (Roy et al., 2005). In the context of single-stage decision making, such a trade off can be quantified using the information value theory (Howard, 1966), which defines it as the reward a decision maker is willing to give away if they could have their uncertainty resolved (sometimes called the expected value of perfect information; EVPI). The give-away amount is the reward difference between a "sophisticated" policy receiving perfect state information and a "naive" policy without such information. Incorporating the sequential nature of decision making and the fact that the agent can only receive a generally noisy observation of the state (i.e., imperfect information), we can obtain a corollary of EVPI for POMDPs (Flaspohler et al., 2020). In this setting, the sophisticated policy is exactly the Bayes optimal policy with value functions defined in (3). The naive policy can be shown to have the following belief state reward and dynamics:

$$R^{open}(b, a) = \sum_{s} b(s)R(s, a), \quad P^{open}(b'|b, a) = \delta(b' - b'(a, b))\,, \tag{4}$$

where $b'(a, b) := b'(s'|b, a) = \sum_s P(s'|s, a)b(s)$. It's clear that the naive policy shares the same reward function as the Bayes optimal policy. However, its belief dynamics misses a counterfactual belief updating operation. Such a belief dynamics has been referred to as *open-loop* in the literature (akin to open-loop controls; Flaspohler et al. 2020) in the sense that the agent planning under this dynamics would choose actions as if it would not be able to observe the environment in the future.

2.3 Active inference

Active inference is an application of the variational principle to perception and action, where intractable Bayesian belief updates (i.e., (2)) are approximated by variational inference (Da Costa et al., 2020). It is well-known that the optimal variational approximation under appropriately chosen family of posterior distributions equals to the exact posterior in (2) (Blei et al., 2017). We will thus assume appropriate choices of variational family and omit suboptimal belief updating in subsequent analyses.

Central to the current discussion is the policy selection objective functions used in active inference, which is its main difference from classic POMDPs. In particular, active inference introduces an objective function called expected free energy (EFE) which, given an initial belief $Q_0(s_0)$ and a finite sequence of actions $a_{0:T-1}$, is defined, in its most popular form, as (Friston et al., 2017):

$$\begin{aligned} &EFE(a_{0:T-1}, Q_0) \\ &\approx \sum_{t=1}^{T} -\underbrace{\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)]}_{\text{Pragmatic value}} - \underbrace{\mathbb{E}_{Q(o_t|a_{0:T-1})}[\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||Q(s_t|a_{0:T-1})]]}_{\text{Epistemic value}}\,. \end{aligned} \tag{5}$$

Here, $Q(s_t|a_{0:T-1}) = \sum_{s_{t-1}} P(s_t|s_{t-1}, a_{t-1})Q(s_{t-1}|a_{0:T-1})$ is the *marginal* state distribution at time step t, $Q(o_t|a_{0:T-1}) = \sum_{s_t} P(o_t|s_t)Q(s_t|a_{0:T-1})$ is the marginal observation distribution, $Q(s_t|o_t, a_{0:T-1}) \propto P(o_t|s_t)Q(s_t|Q_{t-1}, a_{t-1})$ is the future posterior given the marginal of future states as prior and future observations, $\tilde{P}(o_t)$ is a distribution encoding preference over observations, and $\mathbb{KL}$ denotes Kullback Leibler (KL) divergence. (We discuss nuances about this choice in the appendix, which also contains all proofs and derivations).

The first term in (5) is the cross entropy between predicted and preferred observations. Maximizing this term encourages the agent to take actions that realize preference, and thus the label "pragmatic value". The second term is the expected divergence between future prior and posterior beliefs, also known as the expected information gain (IG). Maximizing this term encourages the agent to take informative actions, and thus the label "epistemic value".

It is straightforward to show that the EFE objective, viewed under the belief MDP framework, corresponds to the following reward function:

$$\begin{aligned} R^{EFE}(b,a) &= \mathbb{E}_{P(o'|b,a)}[\log \tilde{P}(o')] + \mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b(s'|o',b,a)||b(s'|b,a)]] \\ &:= \tilde{R}(b,a) + IG(b,a)\,. \end{aligned} \tag{6}$$

Furthermore, it has the same belief dynamics as the open-loop policy defined in (4).

Compared to the Bayes optimal belief MDP in (3), the first reward term $\tilde{R}(b,a)$ is analogous to $R(b,a)$ because it can also be written as a linear combination of the belief. The main difference is in the second term which corresponds to an extra information gain "bonus".

3 EFE approximates Bayes optimal RL policy

The main insight of this paper is that the information gain, or epistemic value, term in EFE contributes to lowering the regret of the open-loop policy, in turn better approximates the Bayes optimal RL policy. We start by introducing our main analysis tool and show the connection between regret, value of information, and information gain. We then present the main results which are the regret bounds for both the open-loop and EFE policies.

3.1 Performance difference in mismatched belief MDPs

We are interested in the regret of the open-loop and EFE policies in a test POMDP environment, for which there exists a Bayes optimal policy given by (3). A special aspect of our setting is that even though both policies will plan using the open-loop belief dynamics in (4), during execution (as is the case in practice) they are allowed to update their beliefs upon observations in the environment. This means that the testing belief dynamics corresponds to the Bayes optimal belief dynamics, and their (internal) planning dynamics are mismatched. We assume the reward functions are correctly specified, i.e., $R^{open}(b,a) = \tilde{R}(b,a) = R(b,a)$. However, EFE has an additional IG term, which means its composite reward is mismatched.

Extending lemma 4.1 of (Vemula et al., 2023) to the setting of mismatched rewards, we can show that the regret or performance gap of a policy π' compared to the oracle or expert policy π is given by:

Lemma 3.1. *(Performance difference in mismatched MDPs) Let π and π' be two policies which are optimal w.r.t. two MDPs M and M'. The two MDPs share the same initial state distribution and discount factor but have different rewards R, R' and dynamics P, P'. Denote $\Delta R(b,a) = R'(b,a) - R(b,a)$. The performance difference between π and π' when both are evaluated in M is given by:*

$$\begin{aligned} J_M(\pi) - J_M(\pi') = &\underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{b\sim d_P^{\pi}}\left[\mathbb{E}_{a\sim\pi(\cdot|b)}[Q_{M'}^{\pi'}(b,a)] - \mathbb{E}_{a'\sim\pi'(\cdot|b')}[Q_{M'}^{\pi'}(b,a)]\right]}_{\text{Policy advantage under expert distribution}} \\ &+ \underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{(b,a)\sim d_P^{\pi'}}\left[\Delta R(b,a) + \gamma\left(\mathbb{E}_{b'\sim P'(\cdot|b,a)}[V_{M'}^{\pi'}(b')] - \mathbb{E}_{b''\sim P(\cdot|b,a)}[V_{M'}^{\pi'}(b'')]\right)\right]}_{\text{Reward-model advantage under own distribution}} \\ &+ \underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{(b,a)\sim d_P^{\pi}}\left[-\Delta R(b,a) + \gamma\left(\mathbb{E}_{b''\sim P(\cdot|b,a)}[V_{M'}^{\pi'}(b'')] - \mathbb{E}_{b'\sim P'(\cdot|b,a)}[V_{M'}^{\pi'}(b')]\right)\right]}_{\text{Reward-model disadvantage under expert distribution}}\,. \end{aligned} \tag{7}$$

Lemma (3.1) shows that other than the difference in how π and π' choose actions as specified by term 1, a major contributor to regret is the difference in their rewards and dynamics models. In fact, Wei et al. (2023) show that the regret scales quadratically (w.r.t. effective planning horizon $\frac{1}{1-\gamma}$) in the squared KL divergence between the two dynamics models.

Interestingly, when π' and π are the open and closed-loop policies, in which case the regret is analogous to the value of information, we can show that the advantage of closed-loop belief dynamics is proportional to information gain:

Proposition 3.2. *Let $R_{max} = \max_{s,a} |R(s,a)|$ and $V^{open}(s)$ be the value function of open-loop policy in open-loop dynamics. The closed-loop model advantage is bounded as follows:*

$$0 \leq \mathbb{E}_{P(b'|b,a)}[V^{open}(b')] - \mathbb{E}_{P^{open}(b''|b,a)}[V^{open}(b'')] \leq \frac{R_{max}}{1-\gamma}\sqrt{2IG(b,a)}\,. \tag{8}$$

3.2 Main result: regret of EFE policy

Proposition 3.2 gives us a clue that if the open-loop policy wants to reduce its performance gap compared to the closed-loop Bayes optimal policy without having to modify its belief dynamics, it could add an information gain bonus to its reward function to cancel out the disadvantage of open-loop belief dynamics (i.e., reward shaping). This corresponds precisely to the EFE belief MDP.

To simplify the comparisons between these policies, we make three assumptions which are formally stated in the appendix. In brief, they assume that the 1) shared reward function between all policies are specified in such a way that the gain in reward under closed-loop belief dynamics outweigh the loss in information gain bonus and 2) the absolute advantage of the EFE policy expected under the expert distribution is no worse than that of the open-loop policy, which is denoted with $\epsilon_{\tilde{\pi}}$.

Then, we show that the performance gaps of the open-loop and EFE policies compared to the Bayes optimal policy are given as follows:

Theorem 3.3. *Let all policies be deployed in POMDP M and all are allowed to update their beliefs according to $b'(o', a, b)$. Let $\epsilon_{IG} = \mathbb{E}_{(b,a)\sim d_P^{\pi}}[IG(b,a)]$ denotes the expected information gain under the Bayes optimal policy's belief-action marginal distribution and let the belief-action marginal induced by both open-loop and EFE policies have bounded density ratio with the Bayes optimal policy $\left\|\frac{d_P^{\tilde{\pi}}(b,a)}{d_P^{\pi}(b,a)}\right\|_{\infty} \leq C$. Under assumptions C.1 and C.2, the performance gap of the open-loop and EFE policies from the optimal policy are bounded as:*

$$\begin{aligned} J_M(\pi) - J_M(\pi^{open}) &\leq \frac{1}{1-\gamma}\epsilon_{\tilde{\pi}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG}\,, \\ J_M(\pi) - J_M(\pi^{EFE}) &\leq \frac{1}{1-\gamma}\epsilon_{\tilde{\pi}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG} - \frac{C+1}{1-\gamma}\epsilon_{IG}\,. \end{aligned} \tag{9}$$

Theorem 3.3 shows that the performance gap of both policies are linear (w.r.t. planning horizon) in the policy advantage and quadratic in the information gain. However, the EFE policy improves over the open-loop policy with a linear increase in information gain.

4 Discussions and conclusion

In this paper, we study the theoretical connection between active inference and reinforcement learning and show that the epistemic value in the EFE objective of active inference can be seen as an approximation to the Bayes optimal RL policy in POMDPs, achieving a linear improvement in regret compared to a naive policy which doesn't take into account the value of information.

Since the EFE objective is non-convex in the belief (because information gain is concave), depending on how the pragmatic value is specified, the value of information may no longer be non-negative and the agent may be distracted by gathering information rather than focusing on collecting task rewards. Thus, pursuing Bayes optimal policies under the EFE objective requires a choice of preference distribution with a suitable "temperature" (in a Boltzmann distribution sense). Conversely, inappropriate choices of temperature may be (and have been) used to explain systematically suboptimal behavior in humans and animals (Engström et al., 2024; Konaka and Naoki, 2023).

Finally, if the EFE objective instead uses closed-loop belief dynamics, as is the case in the more recent "sophisticated" active inference (Friston et al., 2021), the epistemic value may no longer be seen as an approximation. In this case, the EFE objective belongs to a more general class of POMDPs with belief-dependent rewards (Araya et al., 2010), and the preference temperature interpolates reward-seeking and information-seeking, similar to how it interpolates reward-seeking and distribution-matching in MDPs (Da Costa et al., 2023). In this setting, "Bayes optimal" can be interpreted in the dual sense of Bayesian decision theory (i.e., maximizing reward; Berger, 2013) and Bayesian experimental design (i.e., maximizing information gain; Lindley, 1956; MacKay, 1992).

References

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

M. Araya, O. Buffet, V. Thomas, and F. Charpillet. A pomdp extension with belief-dependent rewards. *Advances in neural information processing systems*, 23, 2010.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

T. Champion, H. Bowman, D. Marković, and M. Grześ. Reframing the expected free energy: Four formulations and a unification. *arXiv preprint arXiv:2402.14460*, 2024.

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

L. Da Costa, N. Sajid, T. Parr, K. Friston, and R. Smith. Reward maximization through discrete active inference. *Neural Computation*, 35(5):807–852, 2023.

Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

J. Engström, R. Wei, A. D. McDonald, A. Garcia, M. O'Kelly, and L. Johnson. Resolving uncertainty on the fly: modeling adaptive driving behavior as active inference. *Frontiers in neurorobotics*, 18: 1341750, 2024.

G. Flaspohler, N. A. Roy, and J. W. Fisher III. Belief-dependent macro-action discovery in pomdps using the value of information. *Advances in Neural Information Processing Systems*, 33:11108–11118, 2020.

K. Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

K. Friston, S. Samothrakis, and R. Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106:523–541, 2012.

K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

K. Friston, L. Da Costa, D. Hafner, C. Hesp, and T. Parr. Sophisticated inference. *Neural Computation*, 33(3):713–763, 2021.

K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel. Action and behavior: a free-energy formulation. *Biological cybernetics*, 102:227–260, 2010.

M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, 13:33–94, 2000.

R. A. Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2 (1):22–26, 1966.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Y. Konaka and H. Naoki. Decoding reward–curiosity conflict in decision-making from irrational behaviors. *Nature Computational Science*, 3(5):418–432, 2023.

M. T. Koudahl, W. M. Kouw, and B. de Vries. On epistemics in expected free energy for linear gaussian state space models. *Entropy*, 23(12):1565, 2021.

P. Lanillos, C. Meo, C. Pezzato, A. A. Meera, M. Baioumy, W. Ohata, A. Tschantz, B. Millidge, M. Wisse, C. L. Buckley, et al. Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint arXiv:2112.01871*, 2021.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pages 362–370. Elsevier, 1995.

D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

P. Mazzaglia, T. Verbelen, O. Catal, and B. Dhoedt. The free energy principle for perception and action: A deep learning perspective. *Entropy*, 24(2):301, 2022.

B. Millidge. Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, 96:102348, 2020.

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

T. Parr, G. Pezzulo, and K. J. Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

H. Raiffa and R. Schlaifer. *Applied statistical decision theory*, volume 78. John Wiley & Sons, 2000.

N. Roy, G. Gordon, and S. Thrun. Finding approximate pomdp solutions through belief compression. *Journal of artificial intelligence research*, 23:1–40, 2005.

S. Schwöbel, S. Kiebel, and D. Marković. Active inference, belief propagation, and the bethe approximation. *Neural computation*, 30(9):2530–2567, 2018.

R. Smith, P. Badcock, and K. J. Friston. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. *Psychiatry and Clinical Neurosciences*, 75 (1):3–13, 2021.

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

J. Tomczak and M. Welling. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pages 1214–1223. PMLR, 2018.

A. Tschantz, M. Baltieri, A. K. Seth, and C. L. Buckley. Scaling active inference. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–8. IEEE, 2020.

A. Vemula, Y. Song, A. Singh, D. Bagnell, and S. Choudhury. The virtues of laziness in model-based rl: A unified objective and algorithms. In *International Conference on Machine Learning*, pages 34978–35005. PMLR, 2023.

J. Watson, A. Imohiosen, and J. Peters. Active inference or control as inference? a unifying view. *arXiv preprint arXiv:2010.00262*, 2020.

R. Wei, S. Zeng, C. Li, A. Garcia, A. D. McDonald, and M. Hong. A bayesian approach to robust inverse reinforcement learning. In *Conference on Robot Learning*, pages 2304–2322. PMLR, 2023.

J. Winn, C. M. Bishop, and T. Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.

L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

A Related work

Our work is complementary to prior work examining the relationship between active inference and RL (Millidge et al., 2020; Watson et al., 2020; Da Costa et al., 2023). In (Millidge, 2020) and (Watson et al., 2020), the authors discussed the connections between active inference and control as inference, a popular policy optimization approach in RL inspired by variational inference. However, they consider state space planning as opposed to belief space planning, which (the latter) is a more accurate depiction of active inference (Friston et al., 2021). The closest to our work is (Da Costa et al., 2023) which also considers the optimality of active inference agents in reward-seeking tasks and established the equivalence between the limiting case of EFE and dynamic programming under Bellman's principle of optimality in MDPs. Our work extends (Da Costa et al., 2023) to POMDPs and belief space dynamic programming. Our work is also related to (Schwöbel et al., 2018; Koudahl et al., 2021) which studied the epistemic behavior of active inference agents. Specifically, these works show that the epistemic behavior could disappear in certain dynamical systems or under certain variational approximations. In contrast, we focus on general POMDP environments and study the advantage of using epistemic value.

B Open-loop policy and value of information

Our definition of open-loop policy and value of information is rooted in Howard's information value theory (1966). Here we show how it's obtained.

According to Howard, the expected value of perfect information (EVPI) is defined as:

$$\begin{aligned} EVPI &= EV|PI - EV\,, \\ EV &= \max_a \sum_s b(s)R(s,a)\,, \\ EV|PI &= \sum_s b(s) \max_a R(s,a)\,. \end{aligned} \tag{10}$$

In the POMDP setting, rather than receiving perfect state information, the agent receives a (perfect) observation which provides imperfect information about the state (Raiffa and Schlaifer, 2000). The expected value of perfect observation (EVPO) can be defined as:

$$\begin{aligned} EVPO &= EV|PO - EV\,, \\ EV &= \max_a \sum_s b(s)R(s,a)\,, \\ EV|PO &= \sum_o \sum_s P(o|s)b(s) \max_a R(b(s|o),a)\,. \end{aligned} \tag{11}$$

Adding the sequential nature of POMDPs, we recover the open-loop and Bayes optimal policies as EV and EV|PO, respectively:

$$EV: \quad Q^{open}(b,a) = \sum_s b(s)R(s,a) + \gamma V^{open}(b'(a,b))\,, \tag{12a}$$

$$EV|PO: \quad Q(b,a) = \sum_s b(s)R(s,a) + \gamma \sum_{o'} P(o'|b,a)V(b'(o,a,b))\,. \tag{12b}$$

Note that the definition is the same as (Flaspohler et al., 2020). Here we simply provide more motivation and justification based on Howard (1966) and Raiffa and Schlaifer (2000).

C Assumptions and additional claims

Pragmatic value is linear in the belief This statement appeared in section 3.1. It is straightforward to show:

$$\begin{aligned}\tilde{R}(b,a) &= \mathbb{E}_{P(o'|b,a)}[\log \tilde{P}(o')] \\ &= \sum_s b(s) \sum_{s'} P(s'|s,a) \sum_{o'} P(o'|s') \log \tilde{P}(o') \\ &= \sum_s b(s) \tilde{R}(s,a)\,.\end{aligned} \tag{13}$$

The last line shows that we obtain an analog of state-action reward.

Temperature parameter interpolates reward-seeking and information-seeking This statement appeared in section 4. Let us define the preference distribution as the exponentiated reward multiplied by a negative temperature parameter λ: $\tilde{P}(o) \propto \exp(\lambda \tilde{R}(o))$, then the EFE reward becomes proportional to a weighted combination of reward and information gain:

$$\begin{aligned}\tilde{R}(s,a) &\propto \sum_{s'} P(s'|s,a) \sum_{o'} P(o'|s') \lambda \tilde{R}(o') \\ &= \lambda \tilde{R}(s,a)\,, \\ R^{EFE}(b,a) &\propto \sum_s b(s) \tilde{R}(s,a) + \frac{1}{\lambda} IG(b,a)\,,\end{aligned} \tag{14}$$

where choosing a high $\lambda \to \infty$ corresponds to purely optimizing reward.

Assumption C.1. *(Preference specification) The preference distribution or reward is specified such that the gain in pragmatic value after receiving a new observation is higher than the loss in epistemic value in expectation under the Bayes optimal policy π in closed-loop belief dynamics P:*

$$\mathbb{E}_{(b,a)\sim d_P^{\pi}}\left[\sum_s \left(b(s|o) - b(s)\right) R(s,a)\right] \geq \mathbb{E}_{(b,a)\sim d_P^{\pi}}[IG(b(s),a) - IG(b(s|o),a)]\,. \tag{15}$$

Informally, this assumption ensures the EFE agent does not get distracted by gaining information and still focuses on task relevant behavior. It ensures that the advantage of closed-loop belief dynamics (i.e., (3.2)) under the EFE value function is also non-negative.

Assumption C.2. *(Policy behavior) We make the following assumptions on the behavior of the evaluated policies:*

1. *The absolute advantage of the EFE policy π^{EFE} expected under the Bayes optimal policy's marginal distribution is no worse than that of the open-loop policy π^{open}: $\epsilon_{\tilde{\pi}} = \mathbb{E}_{(b,a)\sim d_P^{\pi}}[|A_P^{\pi^{open}}(b,a)|] \geq \mathbb{E}_{(b,a)\sim d_P^{\pi}}[|A_P^{\pi^{EFE}}(b,a)|]$.*
2. *For both the open-loop policy π^{open} and EFE policy π^{EFE}, it always holds that $IG(b,a) \geq 2$ for any b,a sampled from either their own or the expert policy's marginal distribution.*

Assumption 1 is reasonable because we expect the EFE policy to be more similar to the Bayes optimal policy than the open-loop policy given that the information gain reward encourages information seeking behavior. This enables us to remove policy advantage from the performance gap comparison. Assumption 2 is partly numerically motivated because it allows us to further upper bound the closed-loop model advantage in proposition 3.2 via $\sqrt{2\mathbb{KL}} \leq \mathbb{KL}$ so that the IG reward bonus in EFE can be directly compared with closed-loop model advantage and subtracted from it.

Proposition C.3. *The EFE reward function as defined in (6) is concave in the belief.*

Proof. This statement appears in the section 4.

Recall the EFE reward is defined as:

$$R(b,a) = \mathbb{E}_{P(o'|b,a)}[\log \tilde{P}(o')] + \mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b'(s'|o',b,a)||b'(s'|b,a)]]\,. \tag{16}$$

From (13) we know the first term is linear in the belief b.

The second term can be written as:

$$\begin{aligned}
&\mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b'(s'|o',b,a)||b'(s'|b,a)]] \\
&= \mathbb{E}_{P(o',s'|b,a)}[\log b'(s'|o',b,a) - \log b'(s'|b,a)] \\
&= \mathbb{E}_{P(o',s'|b,a)}[\log b'(s'|b,a) + \log P(o'|s') - \log P(o'|b,a) - \log b'(s'|b,a)] \\
&= \mathbb{E}_{P(o',s'|b,a)}[\log P(o'|s') - \log P(o'|b,a)] \\
&= \mathbb{H}[P(o'|b,a)] - \mathbb{E}_{P(s'|b,a)}[\mathbb{H}[P(o'|s')]] \\
&= -\sum_{o'} P(o'|b,a)\log P(o'|b,a) - \sum_{s} b(s) \sum_{s'} P(s'|s,a)\mathbb{H}[P(o'|s')]\,.
\end{aligned} \tag{17}$$

The second term above is a linear function of the belief.

Applying the definition of convexity to the negative of the first term:

$$\begin{aligned}
&\sum_{o'} P(o'|\lambda b + (1-\lambda)b', a)\log P(o'|\lambda b + (1-\lambda)b', a) \\
&= \sum_{o'}\sum_{s} P(o'|s,a)\,[\lambda b(s) + (1-\lambda)b'(s)]\log\left[\sum_{s}(\lambda b(s)P(o'|s,a) + (1-\lambda)b'(s)P(o'|s,a))\right] \\
&= \sum_{o'}[\lambda P(o'|b,a) + (1-\lambda)P(o'|b,a)]\log\frac{\lambda P(o'|b,a) + (1-\lambda)P(o'|b',a)}{\lambda + (1-\lambda)} \\
&\leq \sum_{o'}\lambda P(o'|b,a)\log P(o'|b,a) + \sum_{o'}(1-\lambda)P(o'|b',a)\log P(o'|b',a)\,,
\end{aligned} \tag{18}$$

where the last line uses the log sum inequality and shows the equation is convex. Thus, the first term is concave and the EFE reward is concave in the belief. □

D Missing proofs

D.1 Proofs for Section 2.3

General definition of EFE We consider the following to be the most general definition of EFE (or full EFE) because it makes the least assumption about $\tilde{P}$ (Friston et al., 2017; Champion et al., 2024).

$$EFE(a_{0:T-1}, Q_0) = \mathbb{E}_{Q(o_{1:T},s_{1:T}|a_{0:T-1})}[\log Q(s_{1:T}|a_{0:T-1}) - \log\tilde{P}(o_{1:T}, s_{1:T})]\,. \tag{19}$$

Derivation of $Q(s_{1:T}|a_{0:T-1})$ in from variational inference We aim to obtain a predictive distribution over future states $s_{1:T}$ given an action sequence $a_{0:T-1}$ using variational inference. Typically, active inference assumes a mean-field factorization of the variational distribution $Q(s_{1:T}|a_{0:T-1}) = \prod_{t=1}^{T} Q(s_t|a_{0:T-1})$. Since there is no observation and thus no likelihood term, the variational free energy $\mathcal{F}$ can be written as:

$$\begin{aligned}
\mathcal{F}(Q) &= \mathbb{E}_{Q(s_{1:T}|a_{0:T-1})}[\log Q(s_{1:T}|a_{0:T-1}) - \log P(s_{1:T}|a_{0:T-1})] \\
&= \mathbb{E}_{Q(s_{1:T}|a_{0:T-1})}\left[\sum_{t=1}^{T}(\log Q(s_t|a_{0:T-1}) - \log P(s_t|s_{t-1}, a_{t-1}))\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{Q(s_{t-1:t}|a_{0:T-1})}[\log Q(s_t|a_{0:T-1}) - \log P(s_t|s_{t-1}, a_{t-1})]\,.
\end{aligned} \tag{20}$$

From (Winn et al., 2005), we know the optimal variational distribution has the form:

$$\begin{aligned}
Q(s_t|a_{0:T-1}) &\propto \exp(\mathbb{E}_{Q(s_{t-1}|a_{0:T-1})}[\log P(s_t|s_{t-1}, a_{t-1})]) \\
&\approx \exp(\log \mathbb{E}_{Q(s_{t-1}|a_{0:T-1})}[P(s_t|s_{t-1}, a_{t-1})]) \\
&= \sum_{s_{t-1}} P(s_t|s_{t-1}, a_{t-1}) Q(s_{t-1}|a_{0:T-1}) \\
&:= Q(s_t|Q_{t-1}, a_{t-1})\,.
\end{aligned} \tag{21}$$

which recovers the definition in section 2.3. The approximation in the second line is due to Jensen's inequality and does not significantly affect our results, because we know from the variational inference literature that the optimal variational distribution must be equal to that of exact inference, which is given by the last line. This also matches the implementation in Pymdp[1], which is one of the main software repositories for active inference.

Active inference and QMDP It is crucial to have a precise definition of the distributions $Q(s_{0:T}|a_{0:T-1})$ and $Q(o_{0:T}, s_{0:T}|a_{0:T-1})$. In the main text, we have specified these as the product of marginal distributions over states and observations. Here, we briefly study the consequences of defining these as the joint distributions:

$$\begin{aligned}
Q(s_{0:T}|a_{0:T-1}) &= b(s_0) \prod_{t=1}^{T} P(s_t|s_{t-1}, a_{t-1})\,, \\
Q(o_{0:T}, s_{0:T}|a_{0:T-1}) &= b(s_0) P(o_0|s_0) \prod_{t=1}^{T} P(s_t|s_{t-1}, a_{t-1}) P(o_t|s_t)\,.
\end{aligned} \tag{22}$$

We start by factorizing the full EFE objective in (19) as:

$$\begin{aligned}
&EFE(a_{0:T-1}) \\
&= \mathbb{E}_{Q(o_{1:T}, s_{1:T}|a_{0:T-1})}[\log Q(s_{1:T}|a_{0:T-1}) - \log \tilde{P}(o_{1:T}, s_{1:T})] \\
&= \mathbb{E}_{Q(o_{1:T}, s_{1:T}|a_{0:T-1})}\left[\sum_{t=1}^{T}\left(\log P(s_t|s_{t-1}, a_{t-1}) - \log \tilde{P}(o_t, s_t)\right)\right] \\
&= \mathbb{E}_{b(s_0)P(s_1|s_0,a_0)P(o_1|s_1)}\Bigg[\log P(s_1|s_0, a_0) - \log \tilde{P}(o_1, s_1) \\
&\qquad + \mathbb{E}_{Q(o_{2:T}, s_{2:T}|s_{0:1}, a_{1:T-1})}\left[\sum_{t=2}^{T}\left(\log P(s_t|s_{t-1}, a_{t-1}) - \log \tilde{P}(o_t, s_t)\right)\right]\Bigg] \\
&= \mathbb{E}_{b(s_0)P(s_1|s_0,a_0)P(o_1|s_1)}\left[\log P(s_1|s_0, a_0) - \log \tilde{P}(o_1, s_1) + EFE(a_{1:T-1})\right] \\
&= \mathbb{E}_{b(s_0)}\left[\mathbb{E}_{P(s_1|s_0,a_0)P(o_1|s_1)}[\log P(s_1|s_0, a_0) - \log \tilde{P}(o_1, s_1)] + \mathbb{E}_{P(s_1|s_0,a_0)}[EFE(a_{1:T-1})]\right]\,.
\end{aligned} \tag{23}$$

This allows us to write down a recursive equation:

$$\begin{aligned}
Q(s_t, a_t) &= \underbrace{\mathbb{E}_{P(s_{t+1}|s_t,a_t)P(o_{t+1}|s_{t+1})}[\log P(s_{t+1}|s_t, a_t) - \log \tilde{P}(o_{t+1}, s_{t+1})]}_{R(s_t,a_t)} + \mathbb{E}_{P(s_{t+1}|s_t,a_t)}[V(s_{t+1})]\,, \\
V(s_t) &= \max_a Q(s_t, a_t)\,,
\end{aligned} \tag{24}$$

and

$$EFE(a_{0:T-1}) = \mathbb{E}_{b(s_0)}[Q(s_0, a_0)]\,. \tag{25}$$

This corresponds to what's known as the QMDP approximation in the POMDP literature (Littman et al., 1995), which is known to overestimate the value of a belief by planning under the implicit assumption that future states will be fully observable (Hauskrecht, 2000).

[1] `https://github.com/infer-actively/pymdp`

EFE bound and choice of preference Despite being the most popular choice of EFE, the pragmatic-epistemic value decomposition (5) is actually a bound on the full EFE defined in (19). To show this, let's consider a single time step since both formulations can be decomposed across time steps. Recall that the pragmatic-epistemic decomposition assumes the following factorization of $\tilde{P}(o, s) = \tilde{P}(o)\tilde{P}(s|o)$. The full EFE can be written as:

$$\begin{aligned}
&EFE_t(a_{0:T-1}) = \mathbb{E}_{Q(o_t,s_t|a_{0:T-1})}[\log Q(s_t|a_{0:T-1}) - \log \tilde{P}(o_t, s_t)] \\
&= -\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)] - \mathbb{E}_{Q(o_t,s_t|a_{0:T-1})}[\log \tilde{P}(s_t|o_t)] + \mathbb{E}_{Q(s_t|a_{0:T-1})}[Q(s_t|a_{0:T-1})] \\
&= -\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)] + \mathbb{E}_{Q(o_t,s_t|a_{0:T-1})}[Q(s_t|o_t, a_{0:T-1})] - \mathbb{E}_{Q(o_t,s_t|a_{0:T-1})}[\log \tilde{P}(s_t|o_t)] \\
&\quad + \mathbb{E}_{Q(s_t|a_{0:T-1})}[Q(s_t|a_{0:T-1})] - \mathbb{E}_{Q(o_t,s_t|a_{0:T-1})}[Q(s_t|o_t, a_{0:T-1})] \\
&= -\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)] + \mathbb{E}_{Q(o_t|a_{0:T-1})}\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||\tilde{P}(s_t|o_t)] \\
&\quad - \mathbb{E}_{Q(o_t|a_{0:T-1})}[\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||Q(s_t|a_{0:T-1})]] \\
&\geq -\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)] - \mathbb{E}_{Q(o_t|a_{0:T-1})}[\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||Q(s_t|a_{0:T-1})]] \, .
\end{aligned} \tag{26}$$

Thus, to keep the bound tight, we could set $\tilde{P}(s|o)$ as:

$$\begin{aligned}
\tilde{P}^*(s|o) &= \arg\min_{\tilde{P}(s|o)} \mathbb{E}_{Q(o_t|a_{0:T-1})}\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||\tilde{P}(s_t|o_t)] \\
&\approx \arg\min_{\tilde{P}(s|o)} \mathbb{E}_{Q(o_t|a_{0:T-1})}\mathbb{KL}[\tilde{P}(s_t|o_t)||Q(s_t|o_t, a_{0:T-1})] \\
&\propto \exp\left(\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log Q(s_t|o_t, a_{0:T-1})]\right) \, ,
\end{aligned} \tag{27}$$

where the approximation in the second line assumes the forward and reverse KL divergences have similar solutions. The result on the last line is sometimes referred to as the aggregate posterior (Tomczak and Welling, 2018). However, since the aggregate posterior depends on the action sequence evaluated, the tightest bound is achieved by an aggregate posterior that updates during each EFE optimization step to ensure that the final aggregate posterior is evaluated under the *optimal* action sequence.

Proposition D.1. *The EFE objective in (5) correspond to a belief MDP with the following reward and dynamics:*

$$R^{EFE}(b, a) = \mathbb{E}_{P(o'|b,a)}[\log \tilde{P}(o')] + \mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b(s'|o', b, a)||b(s'|b, a)]] \tag{28a}$$

$$:= \tilde{R}(b, a) + IG(b, a) \, , \tag{28b}$$

$$P^{open}(b'|b, a) = \delta(b' - b'(a, b)), \text{ where } b'(a, b) := b'(s'|b, a) = \sum_s P(s'|s, a)b(s) \, . \tag{28c}$$

Proof. We proof this by showing that the EFE objective has a Bellman-like decomposition over time steps, by conditioning on the predictive distribution at the previous time step:

$$\begin{aligned}
&EFE(a_{0:T-1}, Q_0) \\
&\approx \sum_{t=1}^{T} -\mathbb{E}_{Q(o_t|a_{0:T-1})}[\log \tilde{P}(o_t)] - \mathbb{E}_{Q(o_t|a_{0:T-1})}[\mathbb{KL}[Q(s_t|o_t, a_{0:T-1})||Q(s_t|a_{0:T-1})]] \\
&= \sum_{t=0}^{T-1} -\mathbb{E}_{Q(o_{t+1}|Q_t,a_t)}[\log \tilde{P}(o_{t+1})] - \mathbb{E}_{Q(o_{t+1}|Q_t,a_t)}[\mathbb{KL}[Q(s_{t+1}|o_{t+1}, Q_t, a_t)||Q(s_{t+1}|Q_t, a_t)]] \\
&= \underbrace{-\mathbb{E}_{Q(o_1|Q_0,a_0)}[\log \tilde{P}(o_1)] - \mathbb{E}_{Q(o_1|Q_0,a_0)}[\mathbb{KL}[Q(s_1|o_1, Q_0, a_0)||Q(s_1|Q_0, a_0)]]}_{R^{EFE}(b,a)} + EFE(a_{1:T-1}, Q_1) \, .
\end{aligned} \tag{29}$$

□

Proposition D.2. *(Active inference policy) The EFE achieved by the optimal action sequence can be equivalently achieved by a time-indexed belief-action policy* $\pi(a_t|Q_t)$.

Proof. We proof this based on Bellman optimality for the full EFE objective in (19), which also holds for the pragmatic-epistemic decomposition.

Starting with the base case:

$$EFE(a_{T-1}, Q_{T-1}) = \mathbb{E}_{Q(o_T, s_T|Q_{T-1}, a_{T-1})}[\log Q(s_T|Q_{T-1}, a_{T-1}) - \log \tilde{P}(o_T, s_T)] \,. \tag{30}$$

It is easy to see that

$$\min_{a_{T-1}} EFE(a_{T-1}, Q_{T-1}) = \max_{\pi_{T-1}} \sum_{a_{T-1}} \pi(a_{T-1}|Q_{T-1}) EFE(a_{T-1}, Q_{T-1}) \,, \tag{31}$$

where the optimal policy is $\pi^*_{T-1}(a_{T-1}|Q_{T-1}) = \delta(a_{T-1} - \arg\min_{\tilde{a}_{T-1}} EFE(\tilde{a}_{T-1}, Q_{T-1}))$.

Applying the identity recursively, we have:

$$\begin{aligned}&\min_{\pi_t} \mathbb{E}_{\pi(a_t|Q_t)}[EFE(a_t, Q_t)] = \min_{\pi_t} \mathbb{E}_{\pi(a_t|Q_t)} \Big\{ \\ &\mathbb{E}_{Q(o_{t+1}, s_{t+1}|Q_t, a_t)}[\log Q(s_{t+1}|Q_t, a_t) - \log \tilde{P}(o_{t+1}, s_{t+1})] + \mathbb{E}_{\pi^*(a_{t+1}|Q_{t+1})}[EFE(a_{t+1}, Q_{t+1})] \Big\} \,.\end{aligned} \tag{32}$$

The optimal policy at each step can be obtained by $\pi(a_t|Q_t) = \delta(a_t - \arg\min_{\tilde{a}_t} EFE(\tilde{a}_t, Q_t))$. □

D.2 Proofs for Section 3.1

D.2.1 Helpful Identities

Proposition D.3. *(Open-loop value function convexity) The open-loop value function as defined in (12a) is piece-wise linear and convex in the beliefs.*

Proof. Recall the definition of the open-loop value function is:

$$Q^{open}(b, a) = \sum_s b(s) R(s, a) + \gamma V^{open}(b'(a, b)) \,. \tag{33}$$

Furthermore, it is a valid belief MDP given the deterministic transition of the belief state defined in (4).

Although this is an infinite horizon value function, due to the contraction mapping property of Bellman equation (Agarwal et al., 2019), it can be approximated arbitrarily close using a finite number of K iterations starting from the base case $Q^{open}_{k=0}(b, a) = \sum_s b(s)R(s, a)$. It is clear the base case value function $V^{open}_{k=0}(b) = \max_{\tilde{a}} Q^{open}_{k=0}(b, \tilde{a})$ is piecewise linear and convex in b.

For iteration $k \in \{1, ..., \infty\}$, we have:

$$Q^{open}_{k+1}(b, a) = \sum_s b(s) R(s, a) + \gamma \max_{a'} Q^{open}_k(b'(a, b), a') \,. \tag{34}$$

The belief update $b'(a, b) = \sum_s P(s'|s, a) b(s)$ is linear and convex in b, making the second term piecewise linear and convex. The first term is also linear and convex. The combination is thus piecewise linear and convex. □

Proposition D.4. *(EVPO non-negativity) Let the expected value of perfect observation for a single stage decision making problem with reward $R(s, a)$, prior belief $b(s)$ and marginal observation distribution $P(o) = \sum_s P(o|s)b(s)$ be defined as:*

$$\begin{aligned} EVPO &= EV|PO - EV \,, \\ EV &= \max_a \sum_s b(s) R(s, a) \,, \\ EV|PO &= \sum_p P(o) \max_a \sum_s b(s|o) R(s, a) \,. \end{aligned} \tag{35}$$

It holds that $EVPO \geq 0$.

Proof. We wish to show:

$$\sum_o P(o) \max_a \sum_s b(s|o)R(s,a) \geq \max_{a'} \sum_s b(s)R(s,a') \,. \tag{36}$$

Let use define $a^*(o) = \arg\max_a \sum_s b(s|o)R(s,a)$, and $a^* = \arg\max_a \sum_s b(s)R(s,a)$ so that we can write the LHS as $\sum_o P(o) \sum_s b(s|o)R(s,a^*(o))$ and the RHS as $\sum_s b(s)R(s,a^*)$.

By definition, we have:

$$\sum_s b(s|o)R(s,a^*(o)) \geq \sum_s b(s|o)R(s,a^*) \,, \tag{37}$$

since $a^*(o)$ is the optimal action taking into consideration of o.

Applying expectation over $P(o)$ to the above inequality, we have:

$$\begin{aligned} \sum_o P(o) \sum_s b(s|o)R(s,a^*(o)) &\geq \sum_o P(o) \sum_s b(s|o)R(s,a^*) \\ &= \sum_s b(s)R(s,a^*) \,, \end{aligned} \tag{38}$$

which completes the proof. □

Proposition D.5. *(EVPO upper bound) Let $R_{max} = \max_{s,a} |R(s,a)|$. The expected value of perfect observation as defined in (35) is upper bounded as follows:*

$$EVPO \leq R_{max} \sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]} \,. \tag{39}$$

Proof. Recall the definition of EVPO is:

$$\begin{aligned} EVPO &= \mathbb{E}_{P(o)}[V(b(s|o))] - V(b(s)) \\ &= \mathbb{E}_{P(o)}\left[\max_{a(o)} \sum_s b(s|o)R(s,a(o))\right] - \max_a \sum_s b(s)R(s,a) \\ &\leq \mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s,a^*(o))\right] - \sum_s b(s)R(s,a^*(o)) \\ &= \mathbb{E}_{P(o)}\left[\sum_s R(s,a^*(o))\,(b(s|o) - b(s))\right] \,, \end{aligned} \tag{40}$$

where we have used $a^*(o) = \arg\max_{a(o)} \sum_s b(s|o)R(s,a(o))$ and the inequality is due to $a^*(o)$ being suboptimal for the second term.

Taking the absolute value of the above EVPO bound, we have:

$$\begin{aligned} |EVPO| &= \left|\mathbb{E}_{P(o)}\left[\sum_s R(s,a^*(o))\,(b(s|o) - b(s))\right]\right| \\ &\overset{(1)}{\leq} \mathbb{E}_{P(o)}\left[\left|\sum_s R(s,a^*(o))\,(b(s|o) - b(s))\right|\right] \\ &\overset{(2)}{\leq} \mathbb{E}_{P(o)}\left[\sum_s |R(s,a^*(o))|\,|b(s|o) - b(s)|\right] \\ &\overset{(3)}{\leq} \|R(\cdot,\cdot)\|_\infty \mathbb{E}_{P(o)}\left[\|b(s|o) - b(s)\|_1\right] \\ &\overset{(4)}{\leq} R_{max} \sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]} \end{aligned} \tag{41}$$

where (1) and (2) are due to Jensen's inequality, (3) is due to Holder's inequality, and (4) is due to Pinsker's inequality. □

Proposition D.6. *(EVPO-POMDP non-negativity) Let $Q^{open}(b,a), V^{open}(b)$ and $Q(b,a), V(b)$ denote the open and closed-loop value functions as defined in (12), it holds that:*

$$Q(b,a) \geq Q^{open}(b,a) \text{ and } V(b) \geq V^{open}(b) \text{ for all } b \in \Delta(\mathcal{S}) \text{ and } a \in \mathcal{A}\,. \tag{42}$$

Proof. Recall the open and closed-loop value functions are defined as:

$$\begin{aligned} Q^{open}(b,a) &= \sum_s b(s)R(s,a) + \gamma V^{open}(b'(a,b)), \quad V^{open}(b) = \max_a Q^{open}(b,a)\,, \\ Q(b,a) &= \sum_s b(s)R(s,a) + \gamma \sum_{o'} P(o'|b,a) V(b'(o',a,b)), \quad V(b) = \max_a Q(b,a)\,. \end{aligned} \tag{43}$$

Although these are infinite horizon value functions, again due to their contraction mapping property (Agarwal et al., 2019), they can be approximated arbitrarily close using a finite number of K iterations starting from the base case $Q_{k=0}(b,a) = \sum_s b(s)R(s,a)$.

Starting with $k = 1$, we have:

$$\begin{aligned} Q_1^{open}(b,a) &= \sum_s b(s)R(s,a) + \gamma V_0^{open}(b'(a,b)), \quad V_0^{open}(b) = \max_a \sum_s b(s)R(s,a)\,, \\ Q_1(b,a) &= \sum_s b(s)R(s,a) + \gamma \sum_{o'} P(o'|b,a) V_0(b'(o',a,b)), \quad V_0(b) = \max_a \sum_s b(s)R(s,a)\,. \end{aligned} \tag{44}$$

Taking the difference between the two value functions and multiply by $\frac{1}{\gamma}$, we have:

$$\begin{aligned} &\frac{1}{\gamma}\left[Q_1(b,a) - Q_1^{open}(b,a)\right] \\ &= \sum_{o'} P(o'|b,a) V_0(b'(o',a,b)) - V_0^{open}(b'(a,b)) \\ &= \sum_{o'} P(o'|b,a) \max_{a^{close}} \sum_s b'(s'|o',a,b) R(s', a^{close}) - \max_{a^{open}} \sum_s b'(s'|a,b) R(s', a^{open}) \\ &= EVPO \geq 0\,, \end{aligned} \tag{45}$$

where the second to last line equals EVPO in proposition D.4 under prior belief $b'(s'|b,a)$ for all $b \in \Delta(s), a \in \mathcal{A}$. Thus it must be non-negative.

Applying the above to the value functions at $k = 1$, we have:

$$\begin{aligned} V_1(b) - V_1^{open}(b) &= \max_{a^{close}} Q_1(b, a^{close}) - \max_{a^{open}} Q_1^{open}(b, a^{open}) \\ &\geq Q_1(b, a^{open*}) - Q_1^{open}(b, a^{open*}) \\ &\geq 0\,, \end{aligned} \tag{46}$$

where we have defined $a^{open*} = \arg\max_{a^{open}} Q_1^{open}(b, a^{open})$.

Now consider $k = 2$, where

$$\begin{aligned} Q_2^{open}(b,a) &= \sum_s b(s)R(s,a) + \gamma V_1^{open}(b'(a,b)), \quad V_1^{open}(b) = \max_a Q_1^{open}(s,a)\,, \\ Q_1(b,a) &= \sum_s b(s)R(s,a) + \gamma \sum_{o'} P(o'|b,a) V_1(b'(o',a,b)), \quad V_1(b) = \max_a Q_1(s,a)\,. \end{aligned} \tag{47}$$

Taking the difference between the two value functions again, we have:

$$
\begin{aligned}
&\frac{1}{\gamma}\left[Q_2(b,a)-Q_2^{open}(b,a)\right] \\
&=\sum_{o'}P(o'|b,a)V_1(b'(o,a,b))-V_1^{open}(b'(a,b)) \\
&=\sum_{o'}P(o'|b,a)\max_{a'^{close}}\left\{\sum_s b'(s|o',a,b)R(s,a'^{close})+\sum_{o''}P(o''|b',a'^{close})V_0(b''(o'',a'^{close},b'))\right\} \\
&\quad-\max_{a'^{open}}\left\{\sum_s b'(s|a,b)R(s,a'^{open})+V_0^{open}(b''(a'^{open},b'))\right\}.
\end{aligned}
\tag{48}
$$

Let $a'^{open*}=\arg\max_{a'^{open}}\left\{\sum_s b'(s|a,b)R(s,a'^{open})+V_0^{open}(b''(a'^{open},b'))\right\}$ and $a'^{close*}=\arg\max_{a'^{close}}\left\{\sum_s b'(s|o',a,b)R(s,a'^{close})+\sum_{o''}P(o''|b',a'^{close})V_0(b''(o'',a'^{close},b'))\right\}$, we have:

$$
\begin{aligned}
&\sum_{o'}P(o'|b,a)\max_{a'^{close}}\left\{\sum_s b'(s|o',a,b)R(s,a'^{close})+\sum_{o''}P(o''|b',a'^{close})V_0(b''(o'',a'^{close},b'))\right\} \\
&\quad-\max_{a'^{open}}\left\{\sum_s b'(s|a,b)R(s,a'^{open})+V_0^{open}(b''(a'^{open},b'))\right\} \\
&\geq\sum_{o'}P(o'|b,a)\max_{a'^{close}}\left\{\sum_s b'(s|o',a,b)R(s,a'^{close})+\sum_{o''}P(o''|b',a'^{open*})V_0(b''(o'',a'^{open*},b'))\right\} \\
&\quad-\left\{\sum_s b'(s|a,b)R(s,a'^{open*})+V_0^{open}(b''(a'^{open*},b'))\right\} \\
&=\underbrace{\sum_{o'}P(o'|b,a)\left\{\max_{a'^{close}}\sum_s b'(s|o',a,b)R(s,a'^{close})-\sum_s b'(s|a,b)R(s,a'^{open*})\right\}}_{EVPO\geq 0} \\
&\quad+\sum_{o'}P(o'|b,a)\underbrace{\left\{\sum_{o''}P(o''|b',a'^{open*})V_0(b''(o'',a'^{open*},b'))-V_0^{open}(b''(a'^{open*},b'))\right\}}_{\geq 0\text{ due to (45)}} \\
&\geq 0\,.
\end{aligned}
\tag{49}
$$

Applying the above to $k\in\{1,...,\infty\}$ recursively, we have:

$$
Q(b,a)\geq Q^{open}(b,a)\text{ and }V(b)\geq V^{open}(b)\,. \tag{50}
$$

□

D.2.2 Main Results of Section 3.1

Lemma D.7. *(Performance difference in mismatched MDPs; restate of lemma 3.1) Let π and π' be two policies which are optimal w.r.t. two MDPs M and M'. The two MDPs share the same initial state distribution and discount factor but have different rewards R, R' and dynamics P, P'. Denote $\Delta R(s,a)=R'(s,a)-R(s,a)$. The performance difference between π and π' when both* *are evaluated in M is given by:*

$$
\begin{aligned}
&J_M(\pi) - J_M(\pi') \\
&= \underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[A_{M'}^{\pi'}(s,a)\right]}_{\text{Advantage under expert distribution}} \\
&\quad + \underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi'}}\left[\Delta R(s,a) + \gamma\left(\mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')] - \mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')]\right)\right]}_{\text{Reward-model advantage under own distribution}} \\
&\quad + \underbrace{\frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[-\Delta R(s,a) + \gamma\left(\mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')] - \mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')]\right)\right]}_{\text{Reward-model disadvantage under expert distribution}}.
\end{aligned}
\tag{51}
$$

Proof. While in the main text we denote states with b to be consistent with the belief MDP notation, here we use s for clarity in the derivation and it introduces no difference in the final result.

Following (Vemula et al., 2023), we expand the performance difference as:

$$
\begin{aligned}
J_M(\pi) - J_M(\pi') &= \mathbb{E}_{\mu(s_0)}[V_M^{\pi}(s_0) - V_M^{\pi'}(s_0)] \\
&= \mathbb{E}_{\mu(s_0)}[V_M^{\pi}(s_0) - V_{M'}^{\pi'}(s_0)] + \mathbb{E}_{\mu(s_0)}[V_{M'}^{\pi'}(s_0) - V_M^{\pi'}(s_0)].
\end{aligned}
\tag{52}
$$

The second term can be expanded as:

$$
\begin{aligned}
&\mathbb{E}_{\mu(s_0)}[V_{M'}^{\pi'}(s_0) - V_M^{\pi'}(s_0)] \\
&= \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi'(\cdot|s_0)}[R'(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - R(s_0,a_0) - \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_M^{\pi'}(s_1)]] \\
&= \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi'(\cdot|s_0)}[\Delta R(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] \\
&\quad + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_M^{\pi'}(s_1)]] \\
&= \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi'(\cdot|s_0)}[\Delta R(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)]] \\
&\quad + \gamma\mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi'(\cdot|s_0),s_1\sim P(\cdot|s_0,a_0)}\underbrace{[V_{M'}^{\pi'}(s_1) - V_M^{\pi'}(s_1)]}_{\text{term a}},
\end{aligned}
\tag{53}
$$

where $\Delta R(s,a) = R'(s,a) - R(s,a)$.

Expanding term a, we arrive at a similar structure to the above:

$$
\begin{aligned}
\text{term a} &= \mathbb{E}_{a_1\sim\pi'(\cdot|s_1)}[R'(s_1,a_1) + \gamma\mathbb{E}_{s_2\sim P'(\cdot|s_1,a_1)}[V_{M'}^{\pi'}(s_2)] - R(s_1,a_1) - \gamma\mathbb{E}_{s_2\sim P(\cdot|s_1,a_1)}[V_M^{\pi'}(s_2)]] \\
&= \mathbb{E}_{a_1\sim\pi'(\cdot|s_1)}[\Delta R(s_1,a_1) + \gamma\mathbb{E}_{s_2\sim P'(\cdot|s_1,a_1)}[V_{M'}^{\pi'}(s_2)] - \gamma\mathbb{E}_{s_2\sim P(\cdot|s_1,a_1)}[V_{M'}^{\pi'}(s_2)]] \\
&\quad + \gamma\mathbb{E}_{a_1\sim\pi'(\cdot|s_1),s_2\sim P(\cdot|s_1,a_1)}\underbrace{[V_{M'}^{\pi'}(s_2) - V_M^{\pi'}(s_2)]}_{\text{term a'}}.
\end{aligned}
\tag{54}
$$

We can thus unroll the last term iteratively and obtain:

$$
\begin{aligned}
&\mathbb{E}_{\mu(s_0)}[V_{M'}^{\pi'}(s_0) - V_M^{\pi'}(s_0)] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(\Delta R(s_t,a_t) + \gamma\mathbb{E}_{s'\sim P'(\cdot|s_t,a_t)}[V_{M'}^{\pi'}(s')] - \gamma\mathbb{E}_{s''\sim P(\cdot|s_t,a_t)}[V_{M'}^{\pi'}(s'')]\right)\right] \\
&= \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi'}}\left[\Delta R(s,a) + \gamma\left(\mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')] - \mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')]\right)\right],
\end{aligned}
\tag{55}
$$

where the expectation in the second line is taken w.r.t. the stochastic process induced by π', P.

We now expand the first term in the performance difference:

$$
\begin{aligned}
&\mathbb{E}_{s_0\sim\mu(\cdot)}[V_M^{\pi}(s_0) - V_{M'}^{\pi'}(s_0)] \\
&= \left(\mathbb{E}_{s_0\sim\mu(\cdot)}[V_M^{\pi}(s_0) - \mathbb{E}_{a_0\sim\pi(\cdot|s_0)}[Q_{M'}^{\pi'}(s_0,a_0)]]\right) + \left(\mathbb{E}_{s_0\sim\mu(\cdot)}[\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}[Q_{M'}^{\pi'}(s_0,a_0)] - V_{M'}^{\pi'}(s_0)]\right) \\
&= \left(\mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[Q_{M'}^{\pi'}(s_0,a_0)] - V_{M'}^{\pi'}(s_0)]\right) \\
&\quad + \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[R(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_M^{\pi}(s_1)]] \\
&\quad - \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[R'(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)]] \\
&= \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[A_{M'}^{\pi'}(s_0,a_0)] \\
&\quad + \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}\Big[-\Delta R(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_M^{\pi}(s_1)] - \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] \\
&\quad + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)]\Big] \\
&= \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[A_{M'}^{\pi'}(s_0,a_0)] \\
&\quad + \mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0)}[-\Delta R(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)] - \gamma\mathbb{E}_{s_1\sim P'(\cdot|s_0,a_0)}[V_{M'}^{\pi'}(s_1)]] \\
&\quad + \gamma\mathbb{E}_{s_0\sim\mu(\cdot),a_0\sim\pi(\cdot|s_0),s_1\sim P(\cdot|s_0,a_0)}\underbrace{[V_M^{\pi}(s_1) - V_{M'}^{\pi'}(s_1)]}_{\text{term b}}\,.
\end{aligned}
\tag{56}
$$

Apply the same unrolling method to term b, we have:

$$
\begin{aligned}
&\mathbb{E}_{s_0\sim\mu(\cdot)}[V_M^{\pi}(s_0) - V_{M'}^{\pi'}(s_0)] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t A_{M'}^{\pi'}(s_t,a_t)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(-\Delta R(s_t,a_t) + \gamma\mathbb{E}_{s''\sim P(\cdot|s_t,a_t)}[V_{M'}^{\pi'}(s'')] - \gamma\mathbb{E}_{s'\sim P'(\cdot|s_t,a_t)}[V_{M'}^{\pi'}(s')]\right)\right] \\
&= \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[A_{M'}^{\pi'}(s,a)\right] \\
&\quad + \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[-\Delta R(s,a) + \gamma\left(\mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')] - \mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')]\right)\right]\,,
\end{aligned}
\tag{57}
$$

where the expectations in the first equality is again taken w.r.t. the stochastic process induced by π, P.

Putting together, we have:

$$
\begin{aligned}
&J_M(\pi) - J_M(\pi') \\
&= \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[A_{M'}^{\pi'}(s,a)\right] \\
&\quad + \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi'}}\left[\Delta R(s,a) + \gamma\left(\mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')] - \mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')]\right)\right] \\
&\quad + \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim d_P^{\pi}}\left[-\Delta R(s,a) + \gamma\left(\mathbb{E}_{s''\sim P(\cdot|s,a)}[V_{M'}^{\pi'}(s'')] - \mathbb{E}_{s'\sim P'(\cdot|s,a)}[V_{M'}^{\pi'}(s')]\right)\right]\,.
\end{aligned}
\tag{58}
$$

□

Proposition D.8. *(Closed-loop model advantage upper bound) Let $R_{max} = \max_{s,a}|R(s,a)|$. The closed-loop model advantage is upper bounded as follows:*

$$
\mathbb{E}_{P(b'|b,a)}[V^{open}(b')] - \mathbb{E}_{P^{open}(b''|b,a)}[V^{open}(b'')] \le \frac{R_{max}}{1-\gamma}\sqrt{2IG(b,a)}\,. \tag{59}
$$

Proof. Recall the closed-loop model advantage is defined as:

$$\mathbb{E}_{P(b'|b,a)}[V(b')] - \mathbb{E}_{P^{open}(b''|b,a)}[V(b'')] = \mathbb{E}_{P(o'|b,a)}[V(b'(s'|o',b,a))] - V(b'(s')) \tag{60}$$

To simplify notation, we will drop the conditioning on b, a in the expectation. This also enables us to remove the "$'$" notation.

We will use a similar method as before where we leverage the contraction mapping property of the value function and start from the base case. It is clear for the base case $k = 0$ where $V(b) = \max_a \sum_s b(s)R(s,a)$, the model advantage is EVPO and thus the upper bound from proposition D.5 applies. To simplify notation, let's denote the upper bound as $C(b)$ since $b(s|o)$ can be calculated from $b(s)$

We now consider $k = 1$:

$$\begin{aligned}
&\mathbb{E}_{P(o)}[V_1(b(s|o))] - V_1(b(s)) \\
&= \mathbb{E}_{P(o)}\left[\max_{a^{close}} \sum_s b(s|o)R(s,a^{close}) + \gamma V_0(b'(a^{close}, b(s|o)))\right] \\
&\quad - \left[\max_{a^{open}} \sum_s b(s)R(s,a^{open}) + \gamma V_0(b'(a^{open}, b(s)))\right] \\
&\le \mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s,a^{close*}) + \gamma V_0(b'(a^{close*}, b(s|o)))\right] \\
&\quad - \left[\sum_s b(s)R(s,a^{close*}) + \gamma V_0(b'(a^{close*}, b(s)))\right] \\
&= \underbrace{\mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s,a^{close*}) - \sum_s b(s)R(s,a^{close*})\right]}_{\text{term a}} \\
&\quad + \gamma \underbrace{\mathbb{E}_{P(o)}\left[V_0(b'(a^{close*}, b(s|o))) - V_0(b'(a^{close*}, b(s)))\right]}_{\text{term b}} .
\end{aligned} \tag{61}$$

Term a is the same as the one in EVPO, thus the upper bound $C(b)$ applies again. In term b, recall the open-loop belief updates are defined as:

$$\begin{aligned}
b'(a, b(s|o)) &= \sum_s P(s'|s,a)b(s|o) := b'(s'|o)\,, \\
b'(a, b(s)) &= \sum_s P(s'|s,a)b(s) := b'(s')\,.
\end{aligned} \tag{62}$$

Due to the convexity of the value functions, we have term b ≥ 0. Furthermore, term b corresponds to EVPO for stage 0 with modified belief updates as defined above. Thus $C(b')$ applies again.

Combining both, we have:

$$\begin{aligned}
&\mathbb{E}_{P(o)}[V_1(b(s|o))] - V_1(b(s)) \\
&\le R_{max}\sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]} + \gamma R_{max}\sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b'(s'|o, a^{close*})||b'(s')]]} \\
&\le R_{max}\sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]} + \gamma R_{max}\sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]}\,,
\end{aligned} \tag{63}$$

where the second inequality is due to data processing inequality.

Applying the above to $k \in \{2, ..., \infty\}$ recursively, we have:

$$\begin{aligned}
\mathbb{E}_{P(o'|b,a)}[V(b'(s'|o'))] - V(b'(s')) &\le R_{max}\sum_{t=0}^{\infty}\gamma^t\sqrt{2\mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b'(s'|o')||b'(s')]]} \\
&= \frac{R_{max}}{1-\gamma}\sqrt{2\mathbb{E}_{P(o'|b,a)}[\mathbb{KL}[b'(s'|o')||b'(s')]]}\,.
\end{aligned} \tag{64}$$

□

D.3 Proofs for Section 3.2

Proposition D.9. *(EFE EVPO upper bound) Let $R_{max} = \max_{s,a} |R(s, a)|$. The expected value of perfect observation as defined in (35) is upper bounded as follows:*

$$EVPO^{EFE} \leq \tilde{R}_{max}\sqrt{2\mathbb{E}_{P(o)}[\mathbb{KL}[b(s|o)||b(s)]]}\,. \tag{65}$$

Proof. Recall the one-step EFE belief reward is:

$$R(b, a) = \sum_s b(s)R(s, a) + IG(b, a)\,, \tag{66}$$

where the reward is defined as $R(s, a) := \tilde{R}(s, a)$ in (6) and $IG(b, a)$ is the information gain. We can thus write EVPO as:

$$\begin{aligned}
&EVPO\\
&= \mathbb{E}_{P(o)}\left[\max_{a(o)}\sum_s b(s|o)R(s, a(o)) + IG(b(s|o), a(o))\right] - \max_a\left[\sum_s b(s)R(s, a) + IG(b(s), a)\right]\\
&\leq \mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s, a^*(o)) + IG(b(s|o), a^*(o))\right] - \left[\sum_s b(s)R(s, a^*(o)) + IG(b(s), a^*(o))\right]\\
&= \mathbb{E}_{P(o)}\left[\sum_s R(s, a^*(o))\,(b(s|o) - b(s))\right] + \underbrace{\mathbb{E}_{P(o)}[IG(b(s|o), a^*(o)) - IG(b(s), a^*(o))]}_{\leq 0}\\
&\leq \mathbb{E}_{P(o)}\left[\sum_s R(s, a^*(o))\,(b(s|o) - b(s))\right]\,,
\end{aligned} \tag{67}$$

where we have used $a^*(o) = \arg\max_{a(o)} \sum_s b(s|o)R(s, a(o))$ and the last inequality is due to IG being a concave function of beliefs. The remaining term is the same as the one in proposition D.5. Thus, applying the result from proposition D.5 we complete the proof. □

Proposition D.10. *(EFE closed-loop model advantage upper bound) Let $R_{max} = \max_{s,a} |R(s, a)|$. The closed-loop model advantage under the EFE value function is upper bounded as follows:*

$$\mathbb{E}_{P(b'|b,a)}[V^{EFE}(b')] - \mathbb{E}_{P^{open}(b''|b,a)}[V^{EFE}(b'')] \leq \frac{R_{max}}{1-\gamma}\sqrt{2IG(b, a)} \tag{68}$$

Proof. Similar to the proof to proposition D.8, we start with the base case which is covered by proposition D.9. To simplify notation, we drop the EFE superscript with the understanding that V^{EFE} is the value function under the EFE belief MDP.

Starting with $k = 1$, we have:

$$
\begin{aligned}
&\mathbb{E}_{P(o)}[V_1(b(s|o))] - V_1(b(s)) \\
&= \mathbb{E}_{P(o)}\left[\max_{a^{close}} \sum_s b(s|o)R(s, a^{close}) + IG(b(s|o), a^{close}) + \gamma V_0(b'(a^{close}, b(s|o)))\right] \\
&\quad - \left[\max_{a^{open}} \sum_s b(s)R(s, a^{open}) + IG(b(s), a^{open}) + \gamma V_0(b'(a^{open}, b(s)))\right] \\
&\leq \mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s, a^{close*}) + IG(b(s|o), a^{close*}) + \gamma V_0(b'(a^{close*}, b(s|o)))\right] \\
&\quad - \left[\sum_s b(s)R(s, a^{close*}) + IG(b(s), a^{close*}) + \gamma V_0(b'(a^{close*}, b(s)))\right] \\
&= \mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s, a^{close*}) - \sum_s b(s)R(s, a^{close*})\right] \\
&\quad + \underbrace{\mathbb{E}_{P(o)}[IG(b(s|o), a^{close*}) - IG(b(s), a^{close*})]}_{\leq 0} \\
&\quad + \gamma \mathbb{E}_{P(o)}\left[V_0(b'(a^{close*}, b(s|o))) - V_0(b'(a^{close*}, b(s)))\right] \\
&\leq \underbrace{\mathbb{E}_{P(o)}\left[\sum_s b(s|o)R(s, a^{close*}) - \sum_s b(s)R(s, a^{close*})\right]}_{\text{term a}} \\
&\quad + \gamma \underbrace{\mathbb{E}_{P(o)}\left[V_0(b'(a^{close*}, b(s|o))) - V_0(b'(a^{close*}, b(s)))\right]}_{\text{term b}}
\end{aligned}
\tag{69}
$$

We arrive at the same form as proposition D.8. While we cannot guarantee term b > 0, the same upper bound holds. The next remark ensures the expected closed-loop model advantage under the EFE reward is non-negative, which provides the motivation for assumption C.1.

Finally, applying the above recursively to $k \in \{2, ..., \infty\}$, we complete the proof. □

Remark D.11. *(Motivation for assumption C.1) To ensure the EFE model advantage expected under the Bayes optimal policy π is non-negative, we need to set the reward such that:*

$$
\mathbb{E}_{(b,a)\sim d_P^{\pi}}\left[\sum_s (b(s|o) - b(s))\, R(s, a)\right] \geq \mathbb{E}_{(b,a)\sim d_P^{\pi}}[IG(b(s), a) - IG(b(s|o), a)]\,, \tag{70}
$$

where d_P^{π} is the marginal distribution induced by the Bayes optimal policy in the closed-loop belief dynamics.

Theorem D.12. *(Open-loop and EFE policy performance gaps; restate of theorem 3.3) Let all policies be deployed in POMDP M and all are allowed to update their beliefs according to $b'(o', a, b)$. Let $\epsilon_{IG} = \mathbb{E}_{(b,a)\sim d_P^{\pi}}[IG(b, a)]$ denotes the expected information gain under the Bayes optimal policy's belief-action marginal distribution and let the belief-action marginal induced by both open-loop and EFE policies have bounded density ratio with the Bayes optimal policy $\left\| \frac{d_P^{\tilde{\pi}}(b,a)}{d_P^{\pi}(b,a)} \right\|_{\infty} \leq C$. Under assumptions C.1 and C.2, the performance gap of the open-loop and EFE policies from the optimal policy are bounded as:*

$$
\begin{aligned}
J_M(\pi) - J_M(\pi^{open}) &\leq \frac{1}{1-\gamma}\epsilon_{\tilde{\pi}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG}\,, \\
J_M(\pi) - J_M(\pi^{EFE}) &\leq \frac{1}{1-\gamma}\epsilon_{\tilde{\pi}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG} - \frac{C+1}{1-\gamma}\epsilon_{IG}\,.
\end{aligned}
\tag{71}
$$

Proof. Let us start by bounding the absolute value of the EFE policy's performance gap:

$$
\begin{aligned}
&|J_M(\pi) - J_M(\pi^{EFE})| \\
&\leq \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}[A^{\pi^{EFE}_{M^{EFE}}}(b,a)]\right| \\
&\quad + \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}\left[-IG(b,a) + \gamma\left(\mathbb{E}_{b''\sim P(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b'')] - \mathbb{E}_{b'\sim P^{open}(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b')]\right)\right]\right| \\
&\quad + \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^{\pi^{EFE}}}\left[IG(b,a) + \gamma\left(\mathbb{E}_{b''\sim P^{open}(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b'')] - \mathbb{E}_{b'\sim P(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b')]\right)\right]\right| .
\end{aligned} \tag{72}
$$

Examining the second term, we have:

$$
\begin{aligned}
&\left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}\left[-IG(b,a) + \gamma\left(\mathbb{E}_{b''\sim P(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b'')] - \mathbb{E}_{b'\sim P^{open}(\cdot|b,a)}[V^{\pi^{EFE}}_{M^{EFE}}(b')]\right)\right]\right| \\
&\leq \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}\left[-IG(b,a) + \frac{\gamma R_{max}}{1-\gamma}\sqrt{2IG(b,a)}\right]\right| \\
&\leq \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}\left[-IG(b,a) + \frac{\gamma R_{max}}{1-\gamma}IG(b,a)\right]\right| \\
&= \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\left|\mathbb{E}_{(b,a)\sim d_P^\pi}[IG(b,a)]\right| \\
&= \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\mathbb{E}_{(b,a)\sim d_P^\pi}[IG(b,a)] .
\end{aligned} \tag{73}
$$

Plugging into the performance gap, we have:

$$
\begin{aligned}
&|J_M(\pi) - J_M(\pi^{EFE})| \\
&\leq \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}[A^{\pi^{EFE}_{M^{EFE}}}(b,a)]\right| \\
&\quad + \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\mathbb{E}_{(b,a)\sim d_P^\pi}[IG(b,a)] + \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\mathbb{E}_{(b,a)\sim d_P^\pi}\left[\left|\frac{d_P^{\pi^{EFE}}(b,a)}{d_P^\pi(b,a)}IG(b,a)\right|\right] \\
&\leq \left|\frac{1}{1-\gamma}\mathbb{E}_{(b,a)\sim d_P^\pi}[A^{\pi^{EFE}_{M^{EFE}}}(b,a)]\right| \\
&\quad + \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\mathbb{E}_{(b,a)\sim d_P^\pi}[IG(b,a)] + \frac{\gamma R_{max} + \gamma - 1}{(1-\gamma)^2}\left\|\frac{d_P^{\pi^{EFE}}(b,a)}{d_P^\pi(b,a)}\right\|_\infty \mathbb{E}_{(b,a)\sim d_P^\pi}\left[|IG(b,a)|\right] \\
&= \frac{1}{1-\gamma}\epsilon_{\pi^{EFE}} + \frac{(C+1)(\gamma R_{max} + \gamma - 1)}{(1-\gamma)^2}\epsilon_{IG} \\
&\leq \frac{1}{1-\gamma}\epsilon_{\pi^{open}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG} - \frac{C+1}{1-\gamma}\epsilon_{IG} .
\end{aligned} \tag{74}
$$

For the open-loop policy which does not have the IG term in the reward, it is easy to see that the performance gap is:

$$
|J_M(\pi) - J_M(\pi^{open})| \leq \frac{1}{1-\gamma}\epsilon_{\pi^{open}} + \frac{(C+1)\gamma R_{max}}{(1-\gamma)^2}\epsilon_{IG} . \tag{75}
$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The goal of this paper is compare the performance gap of the active inference policy and the Bayes optimal RL policy, which is achieved by the body of the paper. In the discussion section we provide partial guidance on preference specification as promised in the abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This paper is theoretical in nature and studies a widely used setting (i.e., POMDPs). We acknowledge not all bounds presented in the paper are the tightest or optimal, however we believe that does not count as limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide all proofs in the appendix and have verified them to the best of our knowledge.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not contain any experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper is theoretical. No data or code is provided.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper is theoretical and does not directly concern area with potential ethical impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper is theoretical and does not directly concern area with potential positive or negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Creators and original owners of assets are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Tis paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.