What Type of Inference is Active Inference?

Wouter W. L. Nuijten[1,2] **Mykola Lukashchuk**[1] **Thijs van de Laar**[1] **Bert de Vries**[1,2]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
[2]Lazy Dynamics, Utrecht, the Netherlands

Abstract

Active inference casts decision-making as inference, with the Expected Free Energy (EFE) unifying goal-directed and information-seeking behavior. Recent work showed that EFE minimization can be written as Variational Free Energy (VFE) minimization on a generative model augmented with epistemic priors. We prove that the VFE of the augmented model can be rewritten as the VFE of the predictive model plus explicit entropy-correction terms, making the EFE contribution transparent. We then show that proper EFE-based planning requires combining these epistemic corrections with a planning correction that turns marginal inference into policy optimization, yielding a full variational characterization of EFE-based planning. This clarifies which corrections are needed for cross-entropy planning and for full EFE-based planning. The same entropy-corrected formulation leads to a detailed message-passing scheme for EFE-based planning together with simpler ablations. Experiments on three grid-world environments show that full EFE-based planning outperforms ablations that omit either the planning correction or the epistemic corrections.

1 INTRODUCTION

Sequential decision-making under uncertainty requires balancing exploitation of current knowledge against exploration to reduce uncertainty. Classical reinforcement learning and optimal control address this through value functions or policy optimization [Sutton and Barto, 2018, Bertsekas, 2012], but typically treat reward maximization and uncertainty reduction as separate objectives.

Planning-as-Inference (PAI) offers an alternative by casting control as probabilistic inference [Attias, 2003, Toussaint, 2009], connecting control to variational inference and message passing [Levine, 2018]. Standard PAI methods optimize objectives such as expected utility or cross-entropy to preferences, but do not include an explicit epistemic drive to reduce environmental uncertainty.

Active inference (AIF) addresses this by minimizing the Expected Free Energy (EFE), unifying instrumental and epistemic objectives [Friston et al., 2015, Da Costa et al., 2020]. Nuijten et al. [2026b] showed that EFE minimization can be reformulated as Variational Free Energy (VFE) minimization on a model augmented with *epistemic priors*. This brings active inference into the variational framework, but leaves open a key distinction: obtaining EFE inside a marginal variational objective is not yet the same as planning over policies. Proper planning additionally requires the planning correction of Lázaro-Gredilla et al. [2024]. This paper makes that separation explicit and derives a message-passing scheme for the combined objective. Our contributions are:

- We show that proper EFE-based planning requires combining two entropy corrections: the planning correction of Lázaro-Gredilla et al. [2024], which turns the expected-utility variational objective into policy optimization, and the epistemic corrections of Nuijten et al. [2026b], which turn marginal VFE minimization into EFE minimization. Together they yield a full variational characterization of EFE-based planning.
- We derive a principled message-passing family for these entropy-corrected objectives. Each added entropy term induces a corresponding channel reparameterization that resolves the circularity of posterior-dependent epistemic priors, and recovers both belief propagation and full active-inference planning within the same derivation.
- We validate the framework on three grid-world environments. They differ in where the value of information gathering lies: in some, sensing actions change the observation model itself; in others, a sensing action's only value is *novelty*, the expected information gain about

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5003–5039.

latent parameters. The alternating heuristic of Nuijten et al. [2026a] finds epistemic actions only in the first case; only the current joint scheme captures novelty.

2 BACKGROUND

2.1 GENERATIVE MODEL FOR SEQUENTIAL DECISION-MAKING

We consider an agent that maintains a generative model predicting future observations, states, and the consequences of actions. Following standard conventions [Levine, 2018, Lázaro-Gredilla et al., 2024], we write this as a rollout model:

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = p(\theta)p(x_0)\prod_{t=1}^{T} p(y_t|x_t, \theta) \cdot p(x_t|x_{t-1}, u_t, \theta)\, p(u_t)\,, \tag{1}$$

where $\boldsymbol{x} = (x_0, \ldots, x_T)$ are latent states, $\boldsymbol{y} = (y_1, \ldots, y_T)$ are observations, $\boldsymbol{u} = (u_1, \ldots, u_T)$ are actions, and θ are unknown model parameters. Here $t = 0$ denotes the current time, and the model predicts a rollout into the future over horizon T. The dynamics $p(x_t|x_{t-1}, u_t, \theta)$ may depend on parameters θ, capturing model uncertainty. Throughout this paper we work in the discrete regime, so all integrals over (y_t, x_t, θ) in what follows reduce to finite sums.

To encode goals, we augment the model with *preference priors* $\hat{p}(x_t)$ and $\hat{p}(y_t)$ over desired states and observations [Levine, 2018]:

$$\hat{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \propto p(\theta)p(x_0)\prod_{t=1}^{T} p(y_t|x_t, \theta)\, p(x_t|x_{t-1}, u_t, \theta) \cdot p(u_t)\, \hat{p}(x_t)\, \hat{p}(y_t)\,. \tag{2}$$

These preference priors can be understood as proportional to exponentiated rewards: $\hat{p}(x) \propto \exp(R(x))$, connecting planning-as-inference to reward maximization [Todorov, 2008]. Together, the rollout model (2) with preferences $\hat{p}(x_t)$ and $\hat{p}(y_t)$ defines our planning problem over horizon T: find a policy $q(u_t|x_{t-1})$ whose induced predicted trajectory agrees with the preferences.

2.2 VARIATIONAL FREE ENERGY

Given a generative model, variational inference approximates the posterior by minimizing the Variational Free Energy over a family of tractable distributions q [Blei et al., 2017]:

$$F_{\hat{p}}[q] = \mathbb{D}_{\mathrm{KL}}\left[q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \| \hat{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)\right]\,. \tag{3}$$

Since all variables are unobserved in the planning setting (they represent future quantities), minimizing $F_{\hat{p}}[q]$ yields beliefs about future trajectories that are consistent with both the dynamics and the preference priors.

2.3 FACTOR GRAPHS AND THE BETHE APPROXIMATION

The generative model (2) factorizes into local terms, which can be represented as a Forney-style factor graph (FFG) [Forney, 2001, Loeliger et al., 2007]. In an FFG, nodes represent factors (probability distributions) and edges represent variables; an edge connects to a node when the variable appears in that factor's scope. We write $\mathcal{E}(a)$ for the set of edges (variables) adjacent to factor node a, and $\mathcal{V}(i)$ for the set of factor nodes adjacent to edge i. The variables in the scope of factor a are denoted $\boldsymbol{s}_a$.

The *Bethe approximation* [Yedidia et al., 2005] exploits this structure by constraining the variational distribution to respect the factorization induced by the graph. Each node a maintains a local belief $q_a(\boldsymbol{s}_a)$ over its adjacent variables $\boldsymbol{s}_a$, and each edge i maintains a singleton belief $q_i(s_i)$. These beliefs must satisfy local consistency constraints:

$$\int q_a(\boldsymbol{s}_a)\, \mathrm{d}\boldsymbol{s}_{a\setminus i} = q_i(s_i) \quad \text{for all } i \in \mathcal{E}(a)\,. \tag{4}$$

Under these constraints, with entropy corrections that prevent double-counting of shared variables, the VFE reduces to the *Bethe Free Energy*:

$$F_{\text{Bethe}}[q] = \sum_{a\in\mathcal{V}} \mathbb{D}_{\mathrm{KL}}\left[q_a(\boldsymbol{s}_a) \| f_a(\boldsymbol{s}_a)\right] + \sum_{i\in\mathcal{E}} (d_i - 1)\, \mathbb{H}\left[q_i(s_i)\right]\,, \tag{5}$$

where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, f_a is the factor at node a, and d_i is the degree (number of connected nodes) of edge i. Minimizing the Bethe Free Energy via message passing yields the belief propagation algorithm; on tree-structured graphs, this recovers exact marginals [Pearl, 1982]. Details are provided in Appendix C.

2.4 EPISTEMIC PRIORS

Standard variational inference does not distinguish between variable types: actions, states, observations, and parameters all enter the VFE symmetrically. Nuijten et al. [2026b] clarified the *epistemic priors* $\tilde{p}(u_t)$, $\tilde{p}(x_t)$, and $\tilde{p}(y_t, x_t)$ that encode which variables are controlled, inferred, or observed. These priors augment the generative model:

$$\tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \propto \hat{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \tilde{p}(u_t)\, \tilde{p}(x_t)\, \tilde{p}(y_t, x_t)\,. \tag{6}$$

Each prior is defined in terms of entropies of conditionals[1]

[1]We write $\mathrm{h}[q(y|x)]$ for the *entropy of the conditional* $q(y|x)$, a function of x, and $\mathbb{H}\left[q(y|x)\right]$ for the *conditional entropy*, a scalar: $\mathbb{H}\left[q(y|x)\right] = \mathbb{E}_{q(x)}\left[\mathrm{h}[q(y|x)]\right]$.

of the variational distribution q:

$$\tilde{p}(u_t) \propto \exp\big(\mathrm{h}[q(x_t, x_{t-1}|u_t)] - \mathrm{h}[q(x_{t-1}|u_t)]\big)\,, \quad (7a)$$

$$\tilde{p}(x_t) \propto \exp\big(\mathbb{E}_{q(\theta|x_t)}\left[-\mathrm{h}[q(y_t|x_t,\theta)]\right]\big)\,, \quad (7b)$$

$$\tilde{p}(y_t, x_t) \propto \exp\big(\mathbb{D}_{\mathrm{KL}}\left[q(\theta|y_t, x_t)\|q(\theta|x_t)\right]\big)\,. \quad (7c)$$

Nuijten et al. [2026b] showed that the VFE of the augmented model $F_{\tilde{p}}[q]$ is an upper bound on the expected EFE. A notable feature is that the epistemic priors depend on the variational distribution q itself, creating a circular dependency that complicates optimization. A central contribution of this paper is to make that circularity explicit as entropy corrections in the objective, rather than leaving it implicit in posterior-dependent priors.

3 RELATED WORK

Planning-as-Inference. The PAI framework casts optimal control as inference in graphical models [Attias, 2003, Toussaint, 2009], connecting control to variational methods and message passing [Levine, 2018]. Closely related formulations include linearly-solvable MDPs [Todorov, 2006], path-integral control [Kappen, 2005], KL control [Kappen et al., 2012], and stochastic optimal control [Rawlik et al., 2012]. A known challenge is *optimistic inference*: conditioning on goals biases posteriors toward trajectories assuming favorable outcomes [Levine, 2018]. This issue was addressed by Lázaro-Gredilla et al. [2024] with an entropy correction that turns the expected-utility variational objective into a proper control objective by penalizing plans that rely on fortuitous state realizations.

Active inference. Active inference minimizes the Expected Free Energy, combining instrumental and epistemic value [Friston et al., 2015, Da Costa et al., 2020, Parr et al., 2022]. Existing methods employ specialized procedures: tree search [Friston et al., 2021], branching [Champion et al., 2022], or dynamic programming [Paul et al., 2024]. The status of the EFE as a variational objective has itself been scrutinized [Millidge et al., 2021], motivating several works that seek to unify EFE with variational inference. In a related direction, Palmieri et al. [2022] combined estimation and control via belief propagation. Building on the Generalized Free Energy [Parr and Friston, 2019], Koudahl et al. [2023] and van de Laar et al. [2024] modified the VFE to include epistemic terms. Most recently, Nuijten et al. [2026b] showed that EFE minimization can be formulated as VFE minimization with epistemic priors, refining the preliminary construction of De Vries et al. [2025], and Nuijten et al. [2026a] implemented this via message passing with alternating updates between the posterior and epistemic priors. A separate, complementary line of work [O'Donoghue et al., 2020, Tarbouriech et al., 2023] casts exploration as posterior inference over value functions, targeting uncertainty in the value function rather than in the model parameters θ. Our contribution is to connect these lines: the epistemic-prior construction provides the EFE correction to a marginal objective, the Lázaro-Gredilla construction provides the planning correction, and their combination yields a principled message-passing formulation of EFE-based planning.

4 ENTROPY CORRECTIONS FOR EFE-BASED PLANNING

We now show that the epistemic priors from Section 2.4 and the planning correction of Lázaro-Gredilla et al. [2024] play different roles. The epistemic priors identify the corrections that transform marginal VFE minimization into EFE minimization. The planning correction turns an expected-utility variational objective into a planning objective over policies. Proper active-inference planning requires both. More broadly, specifying a planning method is a three-way modeling choice: the generative model, the variable-role assignment among controlled, state, parameter, and observed quantities, and the entropy correction selecting the objective. The AIF-specific commitment lives entirely in the last. With no entropy corrections, minimizing the VFE $F_{\hat{p}}[q]$ of the preference-augmented model is simply marginal inference, or in the control setting, KL control [Kappen et al., 2012][2]; different objectives arise by adding entropy corrections to this same baseline. The key question is which corrections are needed for proper EFE-based planning.

4.1 CROSS-ENTROPY PLANNING

Marginal variational inference minimizes a cost over the full q, which lets the joint commit to favorable state realizations that the policy alone cannot produce. Lázaro-Gredilla et al. [2024] showed that turning this into planning, where the extracted policy $q(u_t|x_{t-1})$ actually attains the cost it appears to minimize, requires an entropy correction that penalizes action uncertainty:

$$\sum_{t=1}^{T} \mathbb{H}\left[q(x_{t-1}, u_t)\right] - \mathbb{H}\left[q(x_{t-1})\right] = \sum_{t=1}^{T} \mathbb{H}\left[q(u_t|x_{t-1})\right]\,. \quad (8)$$

See Appendix A.1 for the derivation.

Following the control-as-inference framework [Levine, 2018], rewards can be encoded as preference distributions via $\hat{p}(x) \propto \exp(R(x))$. As shown in Appendix A.3, adding the entropy correction (8) to the VFE transforms the objective into minimizing the *cross-entropy* between the state

[2]Kappen-style tempering $\hat{p}(x) \propto \exp(R(x)/\lambda)$ [Kappen, 2005, Kappen et al., 2012] parameterizes the *generative model* (the preference prior), whereas Table 1 parameterizes the *objective* via entropy corrections; the two axes are orthogonal.

marginals and the preference distribution:

$$\min_q \sum_{t=1}^{T} \mathbb{H}\left[q(x_t), \hat{p}(x_t)\right] + \text{const}\,, \tag{9}$$

where $\mathbb{H}\left[q, \hat{p}\right] = -\mathbb{E}_q\left[\log \hat{p}\right]$ is the cross-entropy. Since $\mathbb{H}\left[q, \hat{p}\right] = -\mathbb{E}_q\left[R(x)\right] + \text{const}$, minimizing cross-entropy is equivalent to maximizing expected reward.

We call this **cross-entropy planning**: the agent maximizes expected reward (equivalently, minimizes cross-entropy to preferences) while committing to a policy.

4.2 EFE AS ENTROPY CORRECTIONS

The epistemic priors introduced in Section 2 augment the generative model with terms that encode variable roles. The VFE of this augmented model can be expressed as the original VFE plus entropy corrections. This rewriting is an exact algebraic identity.

Theorem 1 (Entropy-corrected form of active inference)**.** *The variational objective of Nuijten et al. [2026b] can be written as:*

$$F_{\tilde{p}}[q] = F_{\hat{p}}[q] + \sum_{t=1}^{T} 2\mathbb{H}\left[q(y_t|x_t,\theta)\right] - \mathbb{H}\left[q(x_t|x_{t-1},u_t)\right] - \mathbb{H}\left[q(y_t|x_t)\right]\,. \tag{10}$$

Proof. See Appendix A.2. □

Each prior contributes a specific correction: $\tilde{p}(u_t)$ produces $-\mathbb{H}\left[q(x_t|x_{t-1},u_t)\right]$, $\tilde{p}(x_t)$ produces $+\mathbb{H}\left[q(y_t|x_t,\theta)\right]$, and $\tilde{p}(y_t,x_t)$ contributes a further $+\mathbb{H}\left[q(y_t|x_t,\theta)\right] - \mathbb{H}\left[q(y_t|x_t)\right]$ via the identity $\mathbb{E}_q\left[\mathbb{D}_{\mathrm{KL}}\left[q(\theta|y_t,x_t)\|q(\theta|x_t)\right]\right] = \mathbb{H}\left[q(y_t|x_t)\right] - \mathbb{H}\left[q(y_t|x_t,\theta)\right]$, producing the factor of two. Together these corrections produce EFE minimization inside a marginal variational objective: the $+2\,\mathbb{H}\left[q(y_t|x_t,\theta)\right]$ term pushes beliefs toward state-parameter configurations whose observations are informative, which is what *epistemic* means in AIF. The signs pull in opposite directions: the negative terms favor spreading belief over reachable states and predicted observations, while the positive term favors concentrating it on informative configurations. This tension returns as a min-max structure in the message-passing scheme (Section 5.3). Grouped differently, the observation-side corrections match standard EFE terminology [Da Costa et al., 2020]: one factor of $\mathbb{H}\left[q(y_t|x_t,\theta)\right]$ penalizes *ambiguity*, and the remaining $\mathbb{H}\left[q(y_t|x_t)\right] - \mathbb{H}\left[q(y_t|x_t,\theta)\right]$ enters negatively and rewards *novelty*, the expected information gain about θ. Ambiguity depends only on the observation kernel at a given state; novelty depends on the joint posterior over observations, states, and parameters. This distinction drives the empirical separation in Section 6. By itself, however, (10) does not yet yield EFE-based planning, because it lacks the planning correction (8).

4.3 EFE-BASED PLANNING

The missing step is to combine the marginal-EFE corrections of Theorem 1 with the planning correction of Section 4.1. Adding only the dynamics-side term $-\mathbb{H}\left[q(x_t|x_{t-1},u_t)\right]$ to cross-entropy planning yields **risk-minimizing planning**, which minimizes what Friston et al. [2015] call *risk*; it is a useful intermediate ablation, but not yet full EFE-based planning because it omits the observation-side epistemic corrections.

Appendix B proves that the resulting EFE-based planning objective is

$$\min_q F_{\hat{p}}[q] + \sum_{t=1}^{T} \mathbb{H}\left[q(u_t|x_{t-1})\right] + \sum_{t=1}^{T}\Big(2\mathbb{H}\left[q(y_t|x_t,\theta)\right] - \mathbb{H}\left[q(x_t|x_{t-1},u_t)\right] - \mathbb{H}\left[q(y_t|x_t)\right]\Big)\,. \tag{11}$$

The first sum is the planning correction; the second sum is the EFE correction. Only their combination yields proper **EFE-based planning**.

4.4 COMPARISON OF OBJECTIVES

Table 1 summarizes the progression: the planning correction changes *how* control is posed (cross-entropy planning), the EFE correction changes *what* objective is optimized (marginal EFE), and only their combination yields proper EFE-based planning, the objective implemented in our experiments. The channel reparameterizations required for message passing follow directly from these correction terms (Section 5.1); we turn to that next.

5 MESSAGE PASSING FOR EFE-BASED PLANNING

The full EFE-based planning objective (11) adds four conditional entropy terms to the VFE: one for the policy, one for the dynamics, and two for the observation model. In the Bethe framework a conditional distribution is a ratio of region beliefs, so these terms are not functions of any single coordinate of the optimization problem. We resolve this by introducing auxiliary conditional distributions (*channels*) that promote the corrected conditionals to free variational parameters, yielding a message-passing family that generalizes standard belief propagation.

5.1 CHANNEL REPARAMETERIZATION

The FFG representation (Section 2.3) makes the locality of channel reparameterization explicit: each correction acts

Table 1: Entropy corrections needed to move from baseline variational inference to proper EFE-based planning.

Objective	**Added entropy correction**
Baseline VFE / Marginal inference	0
Cross-entropy planning [Lázaro-Gredilla et al., 2024]	$+\sum_t \mathbb{H}\left[q(u_t\|x_{t-1})\right]$
Risk-minimizing planning	$\text{CE} - \sum_t \mathbb{H}\left[q(x_t\|x_{t-1}, u_t)\right]$
EFE-based planning	$\text{CE} + \sum_t 2\mathbb{H}\left[q(y_t\|x_t, \theta)\right] - \mathbb{H}\left[q(x_t\|x_{t-1}, u_t)\right] - \mathbb{H}\left[q(y_t\|x_t)\right]$

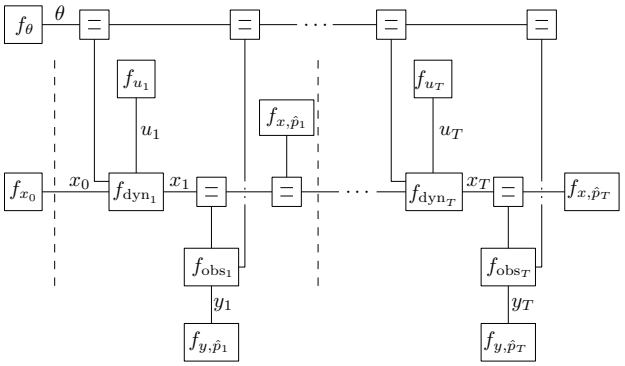


Figure 1: Forney factor graph for the generative model (1). Square nodes are factors; edges are variables. The time slice between the dashed lines is repeated for T timesteps.

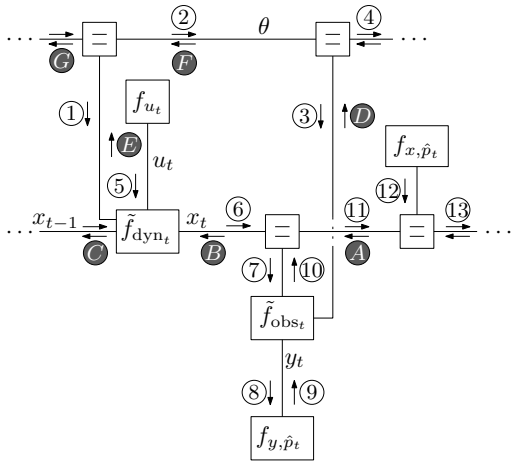


Figure 2: Factor graph for a single time slice of the generative model (1). The observation factor f_{obs_t} and dynamics factor f_{dyn_t} receive kernels $\tilde{f}_{\text{obs}_t}$ and $\tilde{f}_{\text{dyn}_t}$ from (13) under the EFE-based planning objective. Numbered messages are computed in a forward pass over all time slices, lettered messages in a subsequent backward pass.

on a single factor node, while the remainder of the graph is unchanged from standard sum-product (Figure 1). The key identity is the variational characterization of conditional entropy (Gibbs' inequality):

$$\mathbb{H}\left[q(y|x)\right] = \min_r \mathbb{E}_{q(y,x)}\left[-\log r(y|x)\right], \qquad (12)$$

with equality when $r(y|x) = q(y|x)$, so the substitution is exact rather than a bound. We introduce four normalized conditional distributions as *channels*: $r_{u|x,t}(u_t|x_{t-1})$, $r_{x|xu,t}(x_t|x_{t-1}, u_t)$, $r_{y|x\theta,t}(y_t|x_t, \theta)$, and $r_{y|x,t}(y_t|x_t)$, as free variational parameters for each time step t (see Appendix D for formal definitions). Substituting (12) into the corrections of (11) yields a well-posed optimization in which each channel enters the factor it corrects according to the sign of its entropy term: the positive corrections place $r_{u|x,t}$ and $r_{y|x\theta,t}$ (squared) in numerators, while the negative corrections place $r_{x|xu,t}$ and $r_{y|x,t}$ in denominators. This yields the *kernels*:

$$\tilde{f}_{\text{obs}_t}(y_t, x_t, \theta) = \frac{p(y_t|x_t, \theta)\, r^2_{y|x\theta,t}(y_t|x_t, \theta)}{r_{y|x,t}(y_t|x_t)}, \qquad (13a)$$

$$\tilde{f}_{\text{dyn}_t}(x_t, x_{t-1}, \theta, u_t) = \frac{p(x_t|x_{t-1}, \theta, u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1}, u_t)}. \qquad (13b)$$

With these substitutions, the EFE-based planning objective becomes a standard Bethe free energy over the modified factor graph, jointly optimized over beliefs and channels. The kernels (13) replace the original factor functions in the message-passing equations, making the procedure iterative: the channel beliefs r depend on variational beliefs q and vice versa. The proof for $T = 1$ is given in Appendix D. The full scheme comes from the additivity of the Lagrangian and the entropic corrections, see Appendix D.6 for details.

5.2 MESSAGE-PASSING EQUATIONS

Since the modified objective has a Bethe form, the stationarity conditions yield sum-product-style message updates. The only difference from standard belief propagation is that each factor uses its kernel (13) in place of the original.

Each factor a sends to a neighboring factor b the integral of its kernel over incoming messages on adjacent edges:

$$\mu_{jb}(s_j) \propto \int \tilde{f}_a(\boldsymbol{s}_a) \prod_{i \in \mathcal{E}(a)\setminus j} \mu_{ia}(s_i)\, \mathrm{d}\boldsymbol{s}_{a\setminus j}. \qquad (14)$$

For unmodified factors (priors, data likelihoods), $\tilde{f}_a = f_a$.

Singleton beliefs. Singleton beliefs are computed by normalizing the product of colliding messages on an edge,

$$q^*(s_i) \propto \mu_{ia}(s_i)\mu_{ib}(s_i)\,, \tag{15}$$

with $\{a, b\} = \mathcal{V}(i)$ the nodes adjacent to edge i. The full forward-backward schedule is shown in Figure 2.

Region beliefs. Region beliefs are computed by multiplying the factor function with all inbound messages and normalizing,

$$q^*(\boldsymbol{s}_a) \propto \tilde{f}_a(\boldsymbol{s}_a) \prod_{i \in \mathcal{E}(a)} \mu_{ia}(s_i)\,. \tag{16}$$

Channel updates. At the fixed point, each channel recovers the true conditional under its factor belief:

$$r^*_{u|x,t}(u_t|x_{t-1}) = q^*_{u|x,t}(u_t|x_{t-1})\,, \tag{17a}$$

$$r^*_{y|x\theta,t}(y_t|x_t,\theta) = q^*_{y|x\theta,t}(y_t|x_t,\theta)\,, \tag{17b}$$

$$r^*_{y|x,t}(y_t|x_t) = q^*_{y|x,t}(y_t|x_t)\,, \tag{17c}$$

$$r^*_{x|xu,t}(x_t|x_{t-1},u_t) = q^*_{x|xu,t}(x_t|x_{t-1},u_t)\,, \tag{17d}$$

where the beliefs q are conditionals derived from the respective region beliefs around factors f_{obs_t} and f_{dyn_t} at time t. The marginal observation channel $r^*_{y|x,t}$ is obtained by marginalizing θ from the observation factor belief (see Appendix D).

Learning vs. planning. Learning and planning are kept separate. The parameters θ are updated by Bayesian filtering on real observations as they arrive, whereas during planning θ is held fixed at the current belief $q(\theta)$: the backward messages into θ from the (simulated) observation and dynamics factors are not sent. This is what lets the global parameter information gain decompose into the per-step novelty factors $\tilde{p}(y_t, x_t)$ of (7c). Each step's novelty is then scored against a common θ baseline, so it is additive across the horizon. The omitted messages are well-defined; holding θ fixed during planning keeps per-step novelty well-posed.

The same construction yields a family of algorithms: Value Belief Propagation (VBP) [Lázaro-Gredilla et al., 2024] uses only the policy reparameterization, risk-minimizing planning (Table 1) additionally uses the dynamics-side reparameterization, and full EFE-based planning further adds the observation-side reparameterizations. The corresponding VBP derivation is given in Appendix E.

5.3 CONVERGENCE

The kernels (13) contain opposing channel corrections (Section 4): $r_{u|x}$ and $r_{y|x\theta}$ appear in numerators, while $r_{x|xu}$ and $r_{y|x}$ appear in denominators, inducing a min-max structure in the joint optimization. Because each channel reparameterization is a local rewrite of a single kernel, the per-update cost matches standard loopy belief propagation up to the channel updates. Standard BP convergence guarantees, however, do not transfer to this min-max setting, and we apply geometric damping to channels for each update n:

$$r_c^n \propto (r_c^{n-1})^{1-\lambda}\,(r_c^*)^{\lambda}\,, \tag{18}$$

for each channel $c \in \{u|x, x|xu, y|x\theta, y|x\}$. Here $\lambda \in [0, 1]$ is the damping parameter and r_c^* denotes the newly computed channel from (17). We select λ per method and environment from a convergence sweep; at the selected value the channel-based methods reach a stationary VFE plateau, typically within 150 iterations (see Appendix F.5).

Algorithm 1 EFE-Based Planning Message Passing

Input: Generative model factors $\{f_{\mathrm{obs}_t}, f_{\mathrm{dyn}_t}\}_{t=1}^T$, priors $p(\theta), p(x_0), p(u_t)$, goal priors $\hat{p}_x(x_t), \hat{p}_y(y_t)$
Output: Action beliefs $\{q^*_{u_t}(u_t)\}_{t=1}^T$

```
Initialize all messages μ ← 1, channels r_{u|x}, r_{y|xθ}, r_{y|x}, r_{x|xu} ← uniform
repeat
    for t = 1, ..., T do
        Compute sum-product messages (14)
        Update region beliefs (16)
        Update channels (17), (18)
        Update kernels (13)
    end for
    Update singleton beliefs (15)
until convergence
return {q*_{u_t}(u_t)}_{t=1}^T
```

6 EXPERIMENTS

We design experiments to test the behavioral effect of progressively adding the entropy corrections in Table 1. Full experimental details are deferred to Appendix F[3]. All environments use discrete state spaces with exact factor evaluations, isolating the effect of the entropy corrections and channel-augmented schemes from errors introduced by approximate message computation.

6.1 SETUP

Environments. We adapt three classic grid-world environments into epistemic planning benchmarks by treating the environment layout as an unknown parameter θ in the generative model (1). All environments support cardinal movement but differ in how epistemic value arises, which determines which observation-side corrections a method must capture. In Frozen Lake, information gathering changes the observation model itself, so its value is visible as reduced

[3]Code available at `https://github.com/biaslab/UAI-MP-AIF-JAX`

ambiguity. In canonical RockSample and Wumpus World, sensing actions affect only a single observation, so their value is pure novelty: it lies entirely in what the resulting reading reveals about θ.

Frozen Lake [Brockman et al., 2016, Towers et al., 2025]: The agent observes binary "hole/safe" sensors for every cell on the grid, with noise that increases with distance from the agent. A SCAN action switches the agent into a persistent scan mode in which all observations become near-deterministic, at the cost of one time step and a lower prior preference.

RockSample $(4, 3)$ [Smith and Simmons, 2004]: The canonical RockSample benchmark on a 4×4 grid with 3 rocks at known positions and unknown quality (good or bad), defining θ (8 configurations). The agent has one CHECK action per rock, returning a noisy quality reading whose accuracy degrades with distance to that rock, and receives no rock information otherwise. It can SAMPLE a rock for a reward or penalty depending on quality, or EXIT for a fixed reward.

Wumpus World [Russell and Norvig, 1995]: Pit, wumpus, and gold positions define θ (25 configurations). The classic dynamics are simplified to isolate the epistemic challenge: the agent has no orientation or inventory and navigates by cardinal movement. The agent observes noisy breeze, stench, and glitter adjacency signals and has uncertain position. A SCAN action sharpens the adjacency signals for a single step, after which the noise reverts; position channels are unaffected. Even a precise reading only narrows θ: a breeze indicates a nearby pit but not which neighbor, so the agent must triangulate across multiple positions.

Methods. We compare five methods. The first four correspond to message-passing implementations of the entropy-corrected objectives in Section 4.4, with channel configurations as specified in Algorithm 1:

1. **BP**: standard belief propagation, no entropy correction.
2. **VBP**: cross-entropy planning, implemented as the principled channelized scheme from Appendix E.
3. **RM-MP**: risk-minimizing planning (Table 1), using the planning channel together with the dynamics channel; reduces to VBP under deterministic dynamics.
4. **AIF-MP**: full EFE-based planning, using the planning, dynamics, and observation channels (Algorithm 1).
5. **Nuijten-MP** [Nuijten et al., 2026a]: an alternating heuristic that treats the epistemic priors as literal prior factors, recomputed from the current posterior between belief-propagation sweeps, and does not incorporate the novelty prior (7c).

All methods except BP include the planning correction, so the experiments ablate the EFE-side corrections on top of a fixed planning baseline rather than the planning correction itself.

6.2 RESULTS AND DISCUSSION

Table 2 reports performance for all methods across three environments. The results show where each correction matters and where accounting for novelty becomes necessary.

Epistemic actions that change the observation model (Frozen Lake). Both active inference methods dominate ($\sim96\%$ success, overlapping confidence intervals), substantially outperforming all baselines. Both learn to SCAN. Because scan mode persistently changes the observation kernel, reaching scan-mode states already pays off through reduced ambiguity, so the alternating heuristic finds the epistemic action as readily as the joint scheme. RM-MP performs comparably to BP and VBP (overlapping confidence intervals): the dynamics correction is neither beneficial nor harmful in this regime.

Novelty-driven sensing (RockSample and Wumpus World). In canonical RockSample, no action changes the observation model: a CHECK buys a single noisy reading whose only value is the information it carries about rock quality. AIF-MP is the only method that checks and samples, retrieving 98.7% of good rocks for an average reward of 4.01. Every baseline, including Nuijten-MP, walks straight to the exit (reward 1.00, zero retrieval): under the uniform quality prior, sampling an unchecked rock has negative expected reward, so without the novelty term CHECK has no value and EXIT is optimal. Wumpus World repeats this pattern under local readings: even a precise one-step scan does not reveal the global layout, since multiple hazard configurations produce the same breeze and stench patterns, so a scan only narrows θ. AIF-MP again clearly leads (40.7%), while Nuijten-MP performs at the level of VBP (5.0% vs. 5.5%, overlapping confidence intervals): without an exploitable change in the observation model, the alternating heuristic reduces to its planning-only core. RM-MP is the strongest baseline (24.0%): the slip-perturbed dynamics activate its dynamics channel, but without the observation-side corrections it cannot value what a scan reveals about θ. The remaining gap is an objective mismatch, not a scheduling artifact: Nuijten et al. [2026a] treat the epistemic priors as literal prior factors, recomputed *outside* the variational objective between belief-propagation sweeps, and do not incorporate the novelty prior (7c), whereas AIF-MP treats all four channels as variational parameters of a single joint objective with closed-form stationary conditions (17), through which the expected information gain about θ propagates into the plan. Representative trajectories are shown in Appendix F.3.

Synthesis. The alternating heuristic captures half of the observation-side story: it responds to ambiguity, the precision of the observation kernel at reachable states, but not to novelty, the information observations carry about θ. Ambiguity stands in for novelty when an epistemic action persis-

Table 2: Performance across three environments with 95% confidence intervals, averaged over 1000 episodes. Best per metric (non-overlapping CIs) in bold.

	Frozen Lake	**RockSample**		**Wumpus World**
Method	Success (%)	Avg. reward	Retrieval (%)	Success (%)
BP	51.9 [48.8, 55.0]	1.00 [1.00, 1.00]	0.0	1.2 [0.5, 1.9]
VBP	54.5 [51.4, 57.6]	1.00 [1.00, 1.00]	0.0	5.5 [4.1, 6.9]
RM-MP	50.0 [46.9, 53.1]	1.00 [1.00, 1.00]	0.0	24.0 [21.4, 26.6]
Nuijten-MP	**95.6** [94.3, 96.9]	1.00 [1.00, 1.00]	0.0	5.0 [3.6, 6.4]
AIF-MP	**95.9** [94.7, 97.1]	**4.01** [3.90, 4.12]	**98.7**	**40.7** [37.7, 43.7]

tently changes the observation model (Frozen Lake); when sensing actions affect only a single observation, the heuristic collapses to planning-only performance while AIF-MP does not degrade. Across all environments, the planning correction yields modest gains and the dynamics correction helps only under stochastic dynamics (Wumpus World); most of the gap to AIF-MP is explained by the observation-level corrections. Accounting for novelty requires the joint variational treatment; treating the epistemic priors as literal priors and leaving out the novelty prior does not suffice.

7 CONCLUSION

This paper clarifies the variational structure of active inference planning. Theorem 1 shows that the epistemic-prior construction of Nuijten et al. [2026b] admits an explicit entropy-corrected reformulation: relative to baseline VFE minimization, it adds a specific set of entropy corrections that yields marginal EFE minimization, making explicit which terms contribute the epistemic part of the objective. Proper EFE-based planning additionally requires the planning correction of Lázaro-Gredilla et al. [2024], and the combined objective leads directly to a message-passing construction via channel reparameterization. That construction recovers a family of algorithms, including VBP, risk-minimizing planning, and full EFE-based planning (Algorithm 1).

Empirically, the planning and dynamics corrections account for only modest gains, while the observation-side corrections separate along the ambiguity/novelty split: only the joint channelized scheme captures novelty, the expected information gain about model parameters, and sustains performance when sensing actions affect only a single observation.

Limitations and future work. The opposing signs of the entropy corrections induce a min-max structure in the joint optimization over beliefs and channels, which in practice requires damped channel updates for stable convergence (Section 5.3). Standard belief propagation convergence guarantees do not transfer to this setting, and developing convergence theory for the channel-augmented scheme is an open problem; the damping parameter λ also currently requires manual per-environment adjustment. We restrict to discrete state spaces where exact factor evaluations are available; understanding how the channel reparameterization interacts with further factorization constraints on the variational posterior (e.g., mean-field or structured approximations) is an important direction.

Author Contributions

W.W.L. Nuijten and M. Lukashchuk contributed equally to this work. W.W.L. Nuijten developed the entropy decomposition framework. M. Lukashchuk derived the message-passing scheme. Both authors contributed to writing and experiments. T. van de Laar contributed to the conceptualization of the method and supervision. B. de Vries has a supervisory and editorial role.

Acknowledgements

This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), GN Hearing, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023.

References

Hagai Attias. Planning by probabilistic inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 9–16. PMLR, 2003. URL https://proceedings.mlr.press/r4/attias03a.html.

Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*, volume 4. Athena scientific, 2012. URL https://books.google.com/books?hl=en&lr=&id=qVBEEAAAQBAJ&oi=fnd&pg=PR1&dq=Dynamic+Programming+and+Optimal+Control&ots=x0bAav0O5n&sig=s3UxthkdnzR2UpqCUsUsQ7zKgLc.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, June 2016. URL http://arxiv.org/abs/1606.01540.

Théophile Champion, Lancelot Da Costa, Howard Bowman, and Marek Grześ. Branching Time Active Inference: The theory and its generality. *Neural Networks*, 151:295–316, July 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.03.036. URL https://www.sciencedirect.com/science/article/pii/S0893608022001149.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, December 2020. ISSN 0022-2496. doi: 10.1016/j.jmp.2020.102447. URL https://www.sciencedirect.com/science/article/pii/S0022249620300857.

Bert De Vries, Wouter Nuijten, Thijs van de Laar, Wouter Kouw, Sepideh Adamiat, Tim Nisslbeck, Mykola Lukashchuk, Hoang Minh Huu Nguyen, Marco Hidalgo Araya, Raphael Tresor, Thijs Jenneskens, Ivana Nikoloska, Raaja Ganapathy Subramanian, Bart van Erp, Dmitry Bagaev, and Albert Podusenko. Expected Free Energy-based Planning as Variational Inference, April 2025. URL http://arxiv.org/abs/2504.14898.

G. David Forney. Codes on graphs: Normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548, 2001. URL https://ieeexplore.ieee.org/abstract/document/910573/.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, October 2015. ISSN 1758-8928, 1758-8936. doi: 10.1080/17588928.2015.1020053. URL http://www.tandfonline.com/doi/full/10.1080/17588928.2015.1020053.

Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. Sophisticated Inference. *Neural Computation*, 33(3):713–763, March 2021. ISSN 0899-7667. doi: 10.1162/neco_a_01351. URL https://doi.org/10.1162/neco_a_01351.

T. Heskes. Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies. *Journal of Artificial Intelligence Research*, 26:153–190, June 2006. ISSN 1076-9757. doi: 10.1613/jair.1933. URL https://www.jair.org/index.php/jair/article/view/10456.

B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, May 2012. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-012-5278-7. URL http://arxiv.org/abs/0901.0633.

H J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, November 2005. ISSN 1742-5468. doi: 10.1088/1742-5468/2005/11/P11011. URL https://doi.org/10.1088/1742-5468/2005/11/P11011.

Magnus Koudahl, Thijs van de Laar, and Bert de Vries. Realising Synthetic Active Inference Agents, Part I: Epistemic Objectives and Graphical Specification Language, June 2023. URL http://arxiv.org/abs/2306.08014.

Miguel Lázaro-Gredilla, Li Yang Ku, Kevin P. Murphy, and Dileep George. What type of inference is planning? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 116705–116742. Curran Associates, Inc., 2024. doi: 10.52202/079017-3705. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d39e3ae9a11b79691709a7a6e06a63d9-Paper-Conference.pdf.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review, May 2018. URL http://arxiv.org/abs/1805.00909.

Hans-Andrea Loeliger, Justin Dauwels, Junli Hu, Sascha Korl, Li Ping, and Frank R. Kschischang. The Factor Graph Approach to Model-Based Signal Processing. *Proceedings of the IEEE*, 95(6):1295–1322, June 2007. ISSN 0018-9219. doi: 10.1109/JPROC.2007.896497.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the Expected Free Energy? *Neural Computation*, 33(2):447–482, February 2021. ISSN 0899-7667. doi: 10.1162/neco_a_01354. URL https://ieeexplore.ieee.org/abstract/document/9346140.

Wouter W. L. Nuijten, Mykola Lukashchuk, Thijs van de Laar, and Bert de Vries. A Message Passing Realization of Expected Free Energy Minimization. In Mahault Albarracin, David Benrimoh, Christopher L. Buckley, Pablo Lanillos, Riddhi J. Pitliya, Hideaki Shimazaki, Ivilin Peev Stoianov, Tim Verbelen, and Martijn Wisse, editors, *Active Inference*, pages 75–98, Cham, 2026a. Springer Nature Switzerland. ISBN 978-3-032-16955-6. doi: 10.1007/978-3-032-16955-6_5.

Wouter W. L. Nuijten, Thijs van de Laar, and Bert de Vries. Expected free energy-based planning as variational inference. *Transactions on Machine Learning Research*, 2026b. ISSN 2835-8856. URL https://openreview.net/forum?id=Kzm8I1oS1s.

Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xitgHtvS.

Francesco A. N. Palmieri, Krishna R. Pattipati, Giovanni Di Gennaro, Giovanni Fioretti, Francesco Verolla, and Amedeo Buonanno. A Unifying View of Estimation and Control Using Belief Propagation With Application to Path Planning. *IEEE Access*, 10:15193–15216, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3148127.

Thomas Parr and Karl J. Friston. Generalised free energy and active inference. *Biological Cybernetics*, 113(5):495–513, December 2019. ISSN 1432-0770. doi: 10.1007/s00422-019-00805-w. URL https://doi.org/10.1007/s00422-019-00805-w.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press, March 2022. ISBN 978-0-262-36997-8. doi: 10.7551/mitpress/12441.001.0001. URL https://direct.mit.edu/books/book/5299/Active-InferenceThe-Free-Energy-Principle-in-Mind.

Aswin Paul, Noor Sajid, Lancelot Da Costa, and Adeel Razi. On efficient computation in active inference. *Expert Systems with Applications*, 253:124315, November 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.124315. URL http://arxiv.org/abs/2307.00504.

Judea Pearl. Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In *AAAI-82 Proceedings*, pages 133–136, Carnegie Mellon University, Pittsburgh PA, 1982. AAAI Press. URL https://books.google.com/books?hl=nl&lr=&id=e59kEAAAQBAJ&oi=fnd&pg=PA129&ots=qrs53bNhtS&sig=auOq_v4YW1fTTgNOvsmydDN6RPY.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012. URL https://books.google.com/books?hl=en&lr=&id=NOrxCwAAQBAJ&oi=fnd&pg=PA353&dq=On+stochastic+optimal+control+and+reinforcement+learning+by+approximate+inference&ots=dOLjhngmMZ&sig=k04S8PsFkrZCiTHocQskzN34wuk.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 1995.

İsmail Şenöz. *Message Passing Algorithms for Hierarchical Dynamical Models*. Phd Thesis 1 (Research TU/e / Graduation TU/e), Eindhoven University of Technology, Eindhoven, June 2022.

Trey Smith and Reid Simmons. Heuristic search value iteration for pomdps. In *Proceedings of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-04), Arlington, Virginia*, pages 520–527. AUAI Press, 2004.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Jean Tarbouriech, Tor Lattimore, and Brendan O'Donoghue. Probabilistic inference in reinforcement learning done right. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33687–33725. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6a6e010edde1b8f2812f558b67a1974e-Paper-Conference.pdf.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/hash/d806ca13ca3449af72a1ea5aedbed26a-Abstract.html.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292, December 2008. doi: 10.1109/CDC.2008.4739438. URL https://ieeexplore.ieee.org/abstract/document/4739438.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1049–1056, Montreal Quebec Canada, June 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553508. URL https://dl.acm.org/doi/10.1145/1553374.1553508.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d7ff1795e8527f6443371c3933bdb52b-Paper-Datasets_and_Benchmarks_Track.pdf.

Thijs van de Laar, Magnus Koudahl, and Bert de Vries. Realizing synthetic active inference agents, part II: Variational message updates. *Neural Computation*, 37(1): 38–75, 2024.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305, 2008. URL https://www.nowpublishers.com/article/Details/MAL-001.

Jonathan S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.850085.

What Type of Inference is Active Inference? (Supplementary Material)

Wouter W. L. Nuijten[1,2] **Mykola Lukashchuk**[1] **Thijs van de Laar**[1] **Bert de Vries**[1,2]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
[2]Lazy Dynamics, Utrecht, the Netherlands

A ENTROPY CORRECTIONS

A.1 ENTROPY CORRECTION FOR PLANNING

We derive the entropy correction that distinguishes planning-as-inference from marginal inference. Lázaro-Gredilla et al. [2024] formulate planning-as-inference using a "planning entropy" that excludes action variables from the trajectory entropy. Here we show that this formulation is equivalent to adding an entropy correction to the standard VFE, and derive the form of this correction.

A.1.1 The Planning Entropy of Lázaro-Gredilla et al. [2024]

Lázaro-Gredilla et al. [2024] define the planning entropy as:

$$\mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \mathbb{H}_q[x_t|x_{t-1}, u_t]\,, \tag{A.1}$$

where $\mathbb{H}_q[x_t|x_{t-1}, u_t] = -\int q(x_t, x_{t-1}, u_t) \log q(x_t|x_{t-1}, u_t) \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t$ denotes the conditional entropy. This differs from the full trajectory entropy $\mathbb{H}\left[q(\boldsymbol{x}, \boldsymbol{u})\right]$ by excluding the action entropy.

Remark 2 (Notation conventions)**.** *We use u_t to denote the action that leads to state x_t, following the convention in this paper. Lázaro-Gredilla et al. [2024] use a different indexing where u_t leads to x_{t+1}. Additionally, they formulate their objective as a maximization problem (maximizing the variational bound), whereas we minimize the VFE; this flips the sign of entropy terms.*

A.1.2 Derivation of the Entropy Correction

Proposition 3 (Planning entropy decomposition)**.** *The planning entropy* (A.1) *equals the full trajectory entropy plus an entropy correction:*

$$\mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \mathbb{H}_q[x_t|x_{t-1}, u_t] = \mathbb{H}\left[q(\boldsymbol{x}, \boldsymbol{u})\right] + \sum_{t=1}^{T} \mathbb{H}\left[q(x_{t-1})\right] - \mathbb{H}\left[q(x_{t-1}, u_t)\right]\,. \tag{A.2}$$

Proof. Starting from the planning entropy and expanding the conditional entropy:

$$\mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \mathbb{H}_q[x_t|x_{t-1}, u_t]$$

$$= \mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \left(- \iiint q(x_t, x_{t-1}, u_t) \log q(x_t|x_{t-1}, u_t) \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t \right) , \tag{A.3}$$

$$= \mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \left(- \iiint q(x_t, x_{t-1}, u_t) \log \frac{q(x_t, x_{t-1}, u_t)}{q(u_t|x_{t-1}) q(x_{t-1})} \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t \right) . \tag{A.4}$$

Splitting the logarithm:

$$= \mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \left(- \iiint q(x_t, x_{t-1}, u_t) \log \frac{q(x_t, x_{t-1}, u_t)}{q(x_{t-1})} \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t \right.$$

$$\left. + \iiint q(x_t, x_{t-1}, u_t) \log q(u_t|x_{t-1}) \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t \right) , \tag{A.5}$$

$$= \mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \underbrace{\left(- \iiint q(x_t, x_{t-1}, u_t) \log q(x_t, u_t|x_{t-1}) \mathrm{d}x_t \mathrm{d}x_{t-1} \mathrm{d}u_t \right)}_{\mathbb{H}_q[x_t, u_t|x_{t-1}]}$$

$$+ \sum_{t=1}^{T} \iint q(x_{t-1}, u_t) \log \frac{q(x_{t-1}, u_t)}{q(x_{t-1})} \mathrm{d}x_{t-1} \mathrm{d}u_t \, . \tag{A.6}$$

The first sum gives the trajectory entropy by the chain rule:

$$\mathbb{H}\left[q(x_0)\right] + \sum_{t=1}^{T} \mathbb{H}_q[x_t, u_t|x_{t-1}] = \mathbb{H}\left[q(\boldsymbol{x}, \boldsymbol{u})\right] \, . \tag{A.7}$$

The second sum expands as:

$$\sum_{t=1}^{T} \iint q(x_{t-1}, u_t) \log \frac{q(x_{t-1}, u_t)}{q(x_{t-1})} \mathrm{d}x_{t-1} \mathrm{d}u_t = \sum_{t=1}^{T} \underbrace{\iint q(x_{t-1}, u_t) \log q(x_{t-1}, u_t) \mathrm{d}x_{t-1} \mathrm{d}u_t}_{-\mathbb{H}[q(x_{t-1}, u_t)]}$$

$$- \underbrace{\int q(x_{t-1}) \log q(x_{t-1}) \mathrm{d}x_{t-1}}_{-\mathbb{H}[q(x_{t-1})]} , \tag{A.8}$$

$$= \sum_{t=1}^{T} \mathbb{H}\left[q(x_{t-1})\right] - \mathbb{H}\left[q(x_{t-1}, u_t)\right] \, . \tag{A.9}$$

Combining these results proves the proposition. □

A.1.3 Interpretation

Since we minimize the VFE (rather than maximize as in Lázaro-Gredilla et al. [2024]), the planning entropy is *subtracted* from the objective. This means the entropy correction $\sum_t \mathbb{H}\left[q(x_{t-1}, u_t)\right] - \mathbb{H}\left[q(x_{t-1})\right] = \sum_t \mathbb{H}\left[q(u_t|x_{t-1})\right]$ is *added* to the VFE.

Adding this positive correction penalizes action uncertainty: since we minimize the objective, high $\mathbb{H}\left[q(u_t|x_{t-1})\right]$ increases the cost, pushing the agent toward a deterministic policy.

A.2 PROOF OF THEOREM 1

We prove that the VFE of the augmented model (6) decomposes into the original VFE plus entropy correction terms. The proof requires three lemmas, each showing how one epistemic prior contributes to the entropy correction.

Lemma 4 (State epistemic prior contribution)**.** *Let $q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ be a variational distribution over the generative model* (1)*, and let the state epistemic prior be defined as in* (7b)*:*

$$\tilde{p}(x_t) = \exp\big(\mathbb{E}_{q(\theta|x_t)}\left[-\mathrm{h}[q(y_t|x_t,\theta)]\right]\big)\,. \tag{A.10}$$

Then:

$$-\int q(x_t)\log\tilde{p}(x_t)\mathrm{d}x_t = \mathbb{H}\left[q(y_t|x_t,\theta)\right]\,. \tag{A.11}$$

Proof. Substituting the definition of $\tilde{p}(x_t)$ and expanding the conditional entropy:

$$\begin{aligned}
-\int q(x_t)\log\tilde{p}(x_t)\mathrm{d}x_t &= -\int q(x_t)\int q(\theta|x_t)\int q(y_t|x_t,\theta)\log q(y_t|x_t,\theta)\mathrm{d}y_t\mathrm{d}\theta\mathrm{d}x_t && \text{(A.12)}\\
&= -\iiint q(y_t|x_t,\theta)\,q(\theta|x_t)\,q(x_t)\log q(y_t|x_t,\theta)\mathrm{d}y_t\mathrm{d}\theta\mathrm{d}x_t && \text{(A.13)}\\
&= -\iiint q(y_t,x_t,\theta)\log q(y_t|x_t,\theta)\mathrm{d}y_t\mathrm{d}x_t\mathrm{d}\theta && \text{(A.14)}\\
&= \mathbb{H}\left[q(y_t|x_t,\theta)\right]\,. && \text{(A.15)}
\end{aligned}$$

□

Lemma 5 (Action epistemic prior contribution)**.** *Let $q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ be a variational distribution over the generative model* (1)*, and let the action epistemic prior be defined as in* (7a)*:*

$$\tilde{p}(u_t) = \exp\big(\mathrm{h}[q(x_t,x_{t-1}|u_t)] - \mathrm{h}[q(x_{t-1}|u_t)]\big)\,. \tag{A.16}$$

Then:

$$-\int q(u_t)\log\tilde{p}(u_t)\mathrm{d}u_t = -\mathbb{H}\left[q(x_t|x_{t-1},u_t)\right]\,. \tag{A.17}$$

Proof. Substituting the definition of $\tilde{p}(u_t)$:

$$\begin{aligned}
&-\int q(u_t)\log\tilde{p}(u_t)\mathrm{d}u_t\\
&\quad= \int q(u_t)\left(\iint q(x_t,x_{t-1}|u_t)\log q(x_t,x_{t-1}|u_t)\mathrm{d}x_t\mathrm{d}x_{t-1}\right.\\
&\qquad\left.-\int q(x_{t-1}|u_t)\log q(x_{t-1}|u_t)\mathrm{d}x_{t-1}\right)\mathrm{d}u_t && \text{(A.18)}\\
&\quad= \int q(u_t)\left(\iint \frac{q(x_t,x_{t-1},u_t)}{q(u_t)}\log\frac{q(x_t,x_{t-1},u_t)}{q(u_t)}\mathrm{d}x_t\mathrm{d}x_{t-1}\right.\\
&\qquad\left.-\int \frac{q(x_{t-1},u_t)}{q(u_t)}\log\frac{q(x_{t-1},u_t)}{q(u_t)}\mathrm{d}x_{t-1}\right)\mathrm{d}u_t && \text{(A.19)}\\
&\quad= \iiint q(x_t,x_{t-1},u_t)\log\frac{q(x_t,x_{t-1},u_t)}{q(u_t)}\mathrm{d}x_t\mathrm{d}x_{t-1}\mathrm{d}u_t\\
&\qquad-\iint q(x_{t-1},u_t)\log\frac{q(x_{t-1},u_t)}{q(u_t)}\mathrm{d}x_{t-1}\mathrm{d}u_t && \text{(A.20)}\\
&\quad= -\mathbb{H}\left[q(x_t,x_{t-1},u_t)\right] + \mathbb{H}\left[q(u_t)\right] + \mathbb{H}\left[q(x_{t-1},u_t)\right] - \mathbb{H}\left[q(u_t)\right] && \text{(A.21)}\\
&\quad= \mathbb{H}\left[q(x_{t-1},u_t)\right] - \mathbb{H}\left[q(x_t,x_{t-1},u_t)\right] = -\mathbb{H}\left[q(x_t|x_{t-1},u_t)\right]\,. && \text{(A.22)}
\end{aligned}$$

□

Lemma 6 (Observation epistemic prior contribution)**.** *Let $q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ be a variational distribution over the generative model* (1)*, and let the observation epistemic prior be defined as in* (7c)*:*

$$\tilde{p}(y_t, x_t) = \exp\big(\mathbb{D}_{\mathrm{KL}}\left[q(\theta|y_t, x_t) \| q(\theta|x_t)\right]\big) \,. \tag{A.23}$$

Then:

$$-\iint q(y_t, x_t) \log \tilde{p}(y_t, x_t) \mathrm{d}y_t \mathrm{d}x_t = \mathbb{H}\left[q(y_t|x_t, \theta)\right] - \mathbb{H}\left[q(y_t|x_t)\right] \,. \tag{A.24}$$

Proof. Substituting the definition of $\tilde{p}(y_t, x_t)$:

$$\begin{aligned}
&-\iint q(y_t, x_t) \log \tilde{p}(y_t, x_t) \mathrm{d}y_t \mathrm{d}x_t \\
&\quad = -\iint q(y_t, x_t) \left(\int q(\theta|y_t, x_t) \log \frac{q(\theta|y_t, x_t)}{q(\theta|x_t)} \mathrm{d}\theta \right) \mathrm{d}y_t \mathrm{d}x_t && \text{(A.25)} \\
&\quad = -\iint q(y_t, x_t) \left(\int q(\theta|y_t, x_t) \log \frac{q(y_t, x_t, \theta)}{q(y_t, x_t)} - \log \frac{q(x_t, \theta)}{q(x_t)} \mathrm{d}\theta \right) \mathrm{d}y_t \mathrm{d}x_t && \text{(A.26)} \\
&\quad = -\iiint q(y_t, x_t, \theta) \left(\log q(y_t, x_t, \theta) - \log q(y_t, x_t) - \log q(x_t, \theta) + \log q(x_t) \right) \mathrm{d}y_t \mathrm{d}x_t \mathrm{d}\theta && \text{(A.27)} \\
&\quad = \underbrace{-\iiint q(y_t, x_t, \theta) \log q(y_t, x_t, \theta) \mathrm{d}y_t \mathrm{d}x_t \mathrm{d}\theta}_{\mathbb{H}[q(y_t, x_t, \theta)]} \\
&\qquad + \underbrace{\iiint q(y_t, x_t, \theta) \log q(y_t, x_t) \mathrm{d}y_t \mathrm{d}x_t \mathrm{d}\theta}_{-\mathbb{H}[q(y_t, x_t)]} \\
&\qquad + \underbrace{\iiint q(y_t, x_t, \theta) \log q(x_t, \theta) \mathrm{d}y_t \mathrm{d}x_t \mathrm{d}\theta}_{-\mathbb{H}[q(x_t, \theta)]} \\
&\qquad - \underbrace{\iiint q(y_t, x_t, \theta) \log q(x_t) \mathrm{d}y_t \mathrm{d}x_t \mathrm{d}\theta}_{\mathbb{H}[q(x_t)]} && \text{(A.28)} \\
&\quad = \mathbb{H}\left[q(y_t, x_t, \theta)\right] - \mathbb{H}\left[q(y_t, x_t)\right] - \mathbb{H}\left[q(x_t, \theta)\right] + \mathbb{H}\left[q(x_t)\right] && \text{(A.29)} \\
&\quad = \big(\mathbb{H}\left[q(y_t, x_t, \theta)\right] - \mathbb{H}\left[q(x_t, \theta)\right]\big) - \big(\mathbb{H}\left[q(y_t, x_t)\right] - \mathbb{H}\left[q(x_t)\right]\big) && \text{(A.30)} \\
&\quad = \mathbb{H}\left[q(y_t|x_t, \theta)\right] - \mathbb{H}\left[q(y_t|x_t)\right] \,. && \text{(A.31)}
\end{aligned}$$

□

Proof of Theorem 1. The VFE of the augmented model (6) is:

$$F_{\tilde{p}}[q] = \int q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \frac{q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)}{\tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)} \tag{A.32}$$

$$= \int q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \frac{q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)}{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \tilde{p}(x_t)\tilde{p}(u_t)\tilde{p}(y_t, x_t)} \tag{A.33}$$

$$= \underbrace{\int q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \frac{q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)}{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)}}_{F_{\hat{p}}[q]} +$$

$$- \sum_{t=1}^{T} \left(\iiiint q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \tilde{p}(x_t) \mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\mathrm{d}\theta + \iiiint q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \tilde{p}(u_t) \mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\mathrm{d}\theta + \right.$$

$$\left. + \iiiint q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \log \tilde{p}(y_t, x_t) \mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\mathrm{d}\theta \right) \tag{A.34}$$

$$= F_{\hat{p}}[q] + \sum_{t=1}^{T} \left(- \int q(x_t) \log \tilde{p}(x_t) \mathrm{d}x_t - \int q(u_t) \log \tilde{p}(u_t) \mathrm{d}u_t - \iint q(y_t, x_t) \log \tilde{p}(y_t, x_t) \mathrm{d}y_t \mathrm{d}x_t \right) . \tag{A.35}$$

Applying Lemmas 4–6:

$$F_{\tilde{p}}[q] = F_{\hat{p}}[q] + \sum_{t=1}^{T} \Big(\underbrace{\mathbb{H}\left[q(y_t|x_t, \theta)\right]}_{\text{Lemma 4}} \underbrace{- \mathbb{H}\left[q(x_t|x_{t-1}, u_t)\right]}_{\text{Lemma 5}} + \underbrace{\mathbb{H}\left[q(y_t|x_t, \theta)\right] - \mathbb{H}\left[q(y_t|x_t)\right]}_{\text{Lemma 6}} \Big) \tag{A.36}$$

$$= F_{\hat{p}}[q] + \sum_{t=1}^{T} 2\mathbb{H}\left[q(y_t|x_t, \theta)\right] - \mathbb{H}\left[q(x_t|x_{t-1}, u_t)\right] - \mathbb{H}\left[q(y_t|x_t)\right] . \tag{A.37}$$

□

A.3 CROSS-ENTROPY INTERPRETATION

We show that planning-as-inference from Lázaro-Gredilla et al. [2024] is equivalent to minimizing cross-entropy to preference distributions.

A.3.1 Reward as Cross-Entropy

Proposition 7 (Reward as cross-entropy)**.** *For any $\lambda > 0$, define λ-scaled preference distributions $\hat{p}_\lambda(x_t) \propto \exp(\lambda R_t(x_t))$. Then for any distribution $q(\boldsymbol{x}, \boldsymbol{u})$:*

$$\mathbb{E}_{q(\boldsymbol{x}, \boldsymbol{u})} \left[\sum_{t=1}^{T} R_t(x_t) \right] = -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{H}\left[q(x_t), \hat{p}_\lambda(x_t)\right] + \textit{const} . \tag{A.38}$$

Proof. With $\log \hat{p}_\lambda(x_t) = \lambda R_t(x_t) - \log Z_{t,\lambda}$ where $Z_{t,\lambda} = \int \exp(\lambda R_t(x_t)) \mathrm{d}x_t$:

$$\mathbb{H}\left[q(x_t), \hat{p}_\lambda(x_t)\right] = -\mathbb{E}_{q(x_t)} \left[\log \hat{p}_\lambda(x_t)\right] \tag{A.39}$$

$$= -\mathbb{E}_{q(x_t)} \left[\lambda R_t(x_t) - \log Z_{t,\lambda}\right] \tag{A.40}$$

$$= -\lambda \mathbb{E}_{q(x_t)} \left[R_t(x_t)\right] + \log Z_{t,\lambda} . \tag{A.41}$$

Rearranging: $\mathbb{E}_{q(x_t)} \left[R_t(x_t)\right] = -\frac{1}{\lambda}\mathbb{H}\left[q(x_t), \hat{p}_\lambda(x_t)\right] + \frac{1}{\lambda} \log Z_{t,\lambda}$. Summing over t gives the result, where $\frac{1}{\lambda} \sum_t \log Z_{t,\lambda}$ is constant with respect to q. □

A.3.2 Connection to Planning-as-Inference

Lázaro-Gredilla et al. [2024, Theorem 1] show that the best exponential utility

$$F_\lambda^{\text{planning}} = \frac{1}{\lambda} \log \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{u})} \left[\exp\left(\lambda \sum_{t=1}^{T} R_t \right) \right] \tag{A.42}$$

can be expressed as the result of a variational optimization problem whose objective includes the expected sum of rewards $\mathbb{E}_q \left[\sum_t R_t\right]$ as one of its terms.

By Proposition 7 with $\hat{p}_\lambda(x_t) \propto \exp(\lambda R_t(x_t))$, this reward term equals (up to a constant) $-\frac{1}{\lambda} \sum_t \mathbb{H}\left[q(x_t), \hat{p}_\lambda(x_t)\right]$. Since the variational bound in Lázaro-Gredilla et al. [2024, Theorem 1] has an overall $\frac{1}{\lambda}$ scaling, this factor cancels, and maximizing $F_\lambda^{\text{planning}}$ is equivalent to minimizing $\sum_t \mathbb{H}\left[q(x_t), \hat{p}_\lambda(x_t)\right]$ along with the dynamics and entropy terms.

This establishes that with $\hat{p}_\lambda(x_t) \propto \exp(\lambda R_t(x_t))$, the expected utility becomes cross-entropy to preference distributions. Consequently, the inference procedure from Lázaro-Gredilla et al. [2024] minimizes cross-entropy by optimizing its variational objective. Different values of λ yield different preference distributions (corresponding to different risk attitudes), but the underlying mechanism remains cross-entropy minimization.

B EFE-BASED PLANNING INFERENCE

Theorem 8 (EFE-based planning inference)**.** *Consider the augmented model $\tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ from* (6) *and a set of reactive policies $\boldsymbol{\pi} = \{\pi_t(u_t|x_{t-1})\}_{t=1}^T$. Then:*

$$\max_{\boldsymbol{\pi}} \log \sum_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta} \tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \pi_t(u_t|x_{t-1}) = \max_q \left\langle \log \tilde{p} \right\rangle_q + \mathbb{H}\left[q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)\right] - \sum_{t=1}^{T} \mathbb{H}\left[q(u_t|x_{t-1})\right] , \tag{B.1}$$

where $q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ is an arbitrary variational distribution and the optimal policy is $\pi_t^*(u_t|x_{t-1}) = q^*(u_t|x_{t-1})$. Equivalently, up to an additive constant independent of q:

$$\max_{\boldsymbol{\pi}} \log \sum_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta} \tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \pi_t(u_t|x_{t-1}) = -\min_q \; F_{\tilde{p}}[q] + \sum_{t=1}^{T} \mathbb{H}\left[q(u_t|x_{t-1})\right] , \tag{B.2}$$

where $F_{\tilde{p}}[q] = \mathbb{D}_{\mathrm{KL}}\left[q \| \tilde{p}\right]$.

Proof. This is the variational formulation of planning from Lázaro-Gredilla et al. [2024, Appendix A], rewritten in the notation of this paper. To connect our formulation to theirs, note that, up to a normalization constant independent of $\boldsymbol{\pi}$,

$$\begin{aligned} &\sum_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta} \tilde{p}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \pi_t(u_t|x_{t-1}) \\ &\qquad \propto \sum_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta} \rho_{\boldsymbol{\pi}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \prod_{t=1}^{T} \hat{p}(x_t) \hat{p}(y_t) , \end{aligned} \tag{B.3}$$

where

$$\rho_{\boldsymbol{\pi}}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) := p(\theta)\, p(x_0) \prod_{t=1}^{T} p(y_t|x_t, \theta)\, p(x_t|x_{t-1}, u_t, \theta)\, p(u_t)\, \tilde{p}(u_t)\, \tilde{p}(x_t)\, \tilde{p}(y_t, x_t)\, \pi_t(u_t|x_{t-1}) . \tag{B.4}$$

Thus the objective is the log evidence of a preference-weighted rollout model, with $\log \hat{p}(x_t)$ and $\log \hat{p}(y_t)$ playing the role of rewards, exactly as in Appendix A.3. The remaining changes relative to Lázaro-Gredilla et al. [2024] are that: (i) actions are indexed by the state they lead to, so policies are written as $\pi_t(u_t|x_{t-1})$; and (ii) the planning entropy differs from the full trajectory entropy by the correction $\sum_t \mathbb{H}\left[q(u_t|x_{t-1})\right]$ derived in Appendix A.1. The optimal policy satisfies $\pi_t^*(u_t|x_{t-1}) = q^*(u_t|x_{t-1})$ by the same maximization over policies. □

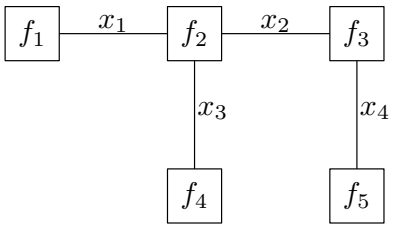


Figure 3: A Forney-style factor graph representing the factorization $p(\boldsymbol{x}) = f_1(x_1)\, f_2(x_1, x_2, x_3)\, f_3(x_2, x_4)\, f_4(x_3)\, f_5(x_4)$. Square nodes denote factors; edges denote variables.

Corollary 9 (Combined entropy correction)**.** *Applying Theorem 1 to expand $F_{\tilde{p}}[q]$, the objective* (B.2) *becomes*

$$\min_q \; F_{\hat{p}}[q] + \underbrace{\sum_{t=1}^{T} 2\mathbb{H}\left[q(y_t|x_t,\theta)\right] - \mathbb{H}\left[q(x_t|x_{t-1},u_t)\right] - \mathbb{H}\left[q(y_t|x_t)\right]}_{\Delta^{\text{AIF}}} + \underbrace{\sum_{t=1}^{T} \mathbb{H}\left[q(u_t|x_{t-1})\right]}_{\Delta^{\text{planning}}} . \tag{B.5}$$

Proof. Direct substitution of Theorem 1 into (B.2). □

C BETHE FREE ENERGY DETAILS

This appendix reviews the standard derivation of the Bethe Free Energy (BFE) from the Variational Free Energy (VFE) on factor graphs, following Yedidia et al. [2005] and Wainwright and Jordan [2008]. The material is collected here for self-containedness and to establish the notation used in Section 2.3 and subsequent appendices.

C.1 FACTORIZED GENERATIVE MODEL ON AN FFG

Consider a generative model that factorizes as

$$p(\boldsymbol{s}) = \prod_{a \in \mathcal{V}} f_a(\boldsymbol{s}_a)\,, \tag{C.1}$$

where each factor f_a depends on a subset of variables $\boldsymbol{s}_a \subseteq \boldsymbol{s}$. This factorization is represented as a Forney-style factor graph (FFG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ [Forney, 2001, Loeliger et al., 2007] (see Figure 3), where $\mathcal{V}$ denotes the set of factor nodes and $\mathcal{E}$ the set of edges (variables). An edge $i \in \mathcal{E}$ connects to a node $a \in \mathcal{V}$ whenever variable s_i appears in the scope of f_a. We write $\mathcal{E}(a)$ for the edges adjacent to node a and $\mathcal{V}(i)$ for the nodes adjacent to edge i. The *degree* of edge i is $d_i = |\mathcal{V}(i)|$, counting the number of factors in which s_i participates.

C.2 THE BETHE VARIATIONAL FAMILY

The Bethe approximation [Yedidia et al., 2005] parameterizes the variational distribution in terms of *local beliefs*: a factor belief $q_a(\boldsymbol{s}_a)$ for each node $a \in \mathcal{V}$ and a singleton belief $q_i(s_i)$ for each edge $i \in \mathcal{E}$. These beliefs define a pseudo-distribution via the *Bethe factorization*:

$$q_{\text{Bethe}}(\boldsymbol{s}) = \frac{\prod_{a \in \mathcal{V}} q_a(\boldsymbol{s}_a)}{\prod_{i \in \mathcal{E}} q_i(s_i)^{d_i - 1}}\,. \tag{C.2}$$

The denominator corrects the over-counting: since each variable s_i appears in d_i factor beliefs, naively multiplying all q_a would count q_i a total of d_i times. Dividing by $q_i^{d_i - 1}$ reduces this to a single effective copy. For a variable that appears in only one factor ($d_i = 1$), no correction is needed.

This parameterization is valid only when the beliefs satisfy *local consistency constraints*:

$$\int q_a(\boldsymbol{s}_a)\, \mathrm{d}\boldsymbol{s}_{a \setminus i} = q_i(s_i) \quad \text{for all } a \in \mathcal{V},\ i \in \mathcal{E}(a)\,, \tag{C.3a}$$

$$\int q_a(\boldsymbol{s}_a)\, \mathrm{d}\boldsymbol{s}_a = 1 \quad \text{for all } a \in \mathcal{V}\,. \tag{C.3b}$$

The marginalization constraint (C.3a) requires that each factor belief, when marginalized over all variables except s_i, agrees with the singleton belief q_i. Together with normalization (C.3b), these constraints define the *local polytope* $\mathcal{L}_\mathcal{G}$.

C.3 DERIVATION: VFE TO BFE

Substituting the model factorization (C.1) and the Bethe factorization (C.2) into the VFE yields the BFE.

Step 1: Expand the log terms. Taking the logarithm of the Bethe factorization:

$$\log q_{\text{Bethe}}(\boldsymbol{s}) = \sum_{a \in \mathcal{V}} \log q_a(\boldsymbol{s}_a) - \sum_{i \in \mathcal{E}} (d_i - 1) \log q_i(s_i) \,. \tag{C.4}$$

Similarly, the log model decomposes as:

$$\log p(\boldsymbol{s}) = \sum_{a \in \mathcal{V}} \log f_a(\boldsymbol{s}_a) \,. \tag{C.5}$$

Step 2: Substitute into the VFE. The VFE is $F[q] = \int q(\boldsymbol{s}) \log \frac{q(\boldsymbol{s})}{p(\boldsymbol{s})} \, \mathrm{d}\boldsymbol{s}$. Substituting (C.4) and (C.5):

$$F[q] = \int q(\boldsymbol{s}) \left[\sum_{a \in \mathcal{V}} \log q_a(\boldsymbol{s}_a) - \sum_{i \in \mathcal{E}} (d_i - 1) \log q_i(s_i) - \sum_{a \in \mathcal{V}} \log f_a(\boldsymbol{s}_a) \right] \mathrm{d}\boldsymbol{s} \,. \tag{C.6}$$

Step 3: Localize expectations using consistency. The key step exploits the local consistency constraints. For any function $g(\boldsymbol{s}_a)$ that depends only on the variables in the scope of factor a:

$$\int q(\boldsymbol{s}) \, g(\boldsymbol{s}_a) \, \mathrm{d}\boldsymbol{s} = \int q_a(\boldsymbol{s}_a) \, g(\boldsymbol{s}_a) \, \mathrm{d}\boldsymbol{s}_a \,, \tag{C.7}$$

since q marginalizes to q_a over $\boldsymbol{s}_a$ by consistency. Similarly, for any function $h(s_i)$ of a single variable:

$$\int q(\boldsymbol{s}) \, h(s_i) \, \mathrm{d}\boldsymbol{s} = \int q_i(s_i) \, h(s_i) \, \mathrm{d}s_i \,. \tag{C.8}$$

Applying these identities to (C.6):

$$F[q] = \sum_{a \in \mathcal{V}} \int q_a(\boldsymbol{s}_a) \log \frac{q_a(\boldsymbol{s}_a)}{f_a(\boldsymbol{s}_a)} \, \mathrm{d}\boldsymbol{s}_a - \sum_{i \in \mathcal{E}} (d_i - 1) \int q_i(s_i) \log q_i(s_i) \, \mathrm{d}s_i \,. \tag{C.9}$$

Step 4: Identify the BFE. Recognizing the KL divergence and entropy, one obtains the *Bethe Free Energy*:

$$F_{\text{Bethe}}[q] = \sum_{a \in \mathcal{V}} \mathbb{D}_{\text{KL}} \left[q_a(\boldsymbol{s}_a) \| f_a(\boldsymbol{s}_a) \right] + \sum_{i \in \mathcal{E}} (d_i - 1) \, \mathbb{H} \left[q_i(s_i) \right] \,, \tag{C.10}$$

which is (5) in the main text. The first sum penalizes each factor belief for deviating from its corresponding factor, while the second sum adds back the entropy of shared variables to correct for the over-counting inherent in the Bethe factorization.

C.4 CONSTRAINED OPTIMIZATION AND BELIEF PROPAGATION

The BFE gives rise to the constrained optimization problem

$$\min_{\{q_a, q_i\} \in \mathcal{L}_\mathcal{G}} F_{\text{Bethe}}[q] \,, \tag{C.11}$$

where $\mathcal{L}_\mathcal{G}$ is the local polytope defined by the constraints (C.3). Yedidia et al. [2005] showed that the stationary points of this constrained problem correspond exactly to the fixed points of the *belief propagation* (BP) algorithm.

On *tree-structured* graphs, the BFE equals the exact VFE, and BP converges to the exact posterior marginals [Pearl, 1982]. In this case, the Bethe factorization (C.2) is an exact representation of the global posterior, and the local consistency constraints are sufficient to characterize it.

On graphs with *loops*, the BFE is an approximation: the Bethe factorization does not generally correspond to a valid probability distribution, and BP may not converge. Nevertheless, when BP does converge, its fixed points remain stationary points of the BFE [Yedidia et al., 2005, Heskes, 2006].

Remark 10. *The Bethe approximation is one member of a broader family. Manipulating the local constraints (for instance, imposing full factorization $q(\boldsymbol{s}) = \prod_i q_i(s_i)$) recovers the mean-field approximation. Other constraint choices yield generalizations such as the Kikuchi free energy [Yedidia et al., 2005]. Different constraint manipulations on the factor graph unify variational message passing, belief propagation, and expectation propagation within a single framework [Şenöz, 2022].*

In models with shared parameters, such as a temporal chain where θ appears in both the dynamics and observation factors at every time step, the resulting loops make the BFE an approximation rather than an exact decomposition. Entropic corrections to the BFE yield modified objectives whose stationarity conditions are derived in Appendix D.

D DETAILED MESSAGE PASSING DERIVATION FOR THE COMBINED OBJECTIVE

This appendix gives a self-contained $T = 1$ derivation of the message-passing scheme for the combined objective. The objective includes both the cross-entropy planning correction and the entropy corrections required for planning-as-inference in active inference. Its message-passing structure is organized around four channels, with $r_{u|x}(u_1|x_0)$ appearing in the numerator of the dynamics kernel.

D.1 MODEL, COORDINATES, AND CONSTRAINTS

We consider the biased generative model

$$p(y_1, x_1, x_0, \theta, u_1) = p(\theta)\, p(x_0)\, p(u_1)\, p(x_1|x_0, \theta, u_1)\, p(y_1|x_1, \theta)\, \hat{p}_x(x_1)\, \hat{p}_y(y_1). \tag{D.1}$$

The corresponding non-singleton factors are the observation factor $f_{\mathrm{obs}}(y_1, x_1, \theta) = p(y_1|x_1, \theta)$ and the dynamics factor $f_{\mathrm{dyn}}(x_1, x_0, \theta, u_1) = p(x_1|x_0, \theta, u_1)$, together with singleton prior and goal factors on θ, x_0, u_1, x_1, and y_1.

The Bethe coordinates consist of the factor beliefs

$$q_{\mathrm{obs}}(y_1, x_1, \theta), \qquad q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1), \tag{D.2}$$

the singleton beliefs are

$$q_\theta(\theta), \quad q_{x_0}(x_0), \quad q_{x_1}(x_1), \quad q_{y_1}(y_1), \quad q_{u_1}(u_1), \tag{D.3}$$

and the local-polytope constraints consist of factor normalization,

$$\int q_{\mathrm{obs}}(y_1, x_1, \theta)\, \mathrm{d}y_1\, \mathrm{d}x_1\, \mathrm{d}\theta = 1, \tag{D.4a}$$

$$\int q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\, \mathrm{d}x_1\, \mathrm{d}x_0\, \mathrm{d}\theta\, \mathrm{d}u_1 = 1, \tag{D.4b}$$

and factor-to-singleton consistency,

$$\int q_{\mathrm{obs}}(y_1, x_1, \theta)\, \mathrm{d}x_1\, \mathrm{d}\theta = q_{y_1}(y_1), \tag{D.5a}$$

$$\int q_{\mathrm{obs}}(y_1, x_1, \theta)\, \mathrm{d}y_1\, \mathrm{d}\theta = q_{x_1}(x_1), \tag{D.5b}$$

$$\int q_{\mathrm{obs}}(y_1, x_1, \theta)\, \mathrm{d}y_1\, \mathrm{d}x_1 = q_\theta(\theta), \tag{D.5c}$$

$$\int q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\, \mathrm{d}x_0\, \mathrm{d}\theta\, \mathrm{d}u_1 = q_{x_1}(x_1), \tag{D.5d}$$

$$\int q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\, \mathrm{d}x_1\, \mathrm{d}\theta\, \mathrm{d}u_1 = q_{x_0}(x_0), \tag{D.5e}$$

$$\int q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\, \mathrm{d}x_1\, \mathrm{d}x_0\, \mathrm{d}u_1 = q_\theta(\theta), \tag{D.5f}$$

$$\int q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\, \mathrm{d}x_1\, \mathrm{d}x_0\, \mathrm{d}\theta = q_{u_1}(u_1). \tag{D.5g}$$

The combined objective uses four normalized channels:

$$r_{y|x\theta}(y_1|x_1,\theta), \qquad r_{y|x}(y_1|x_1), \qquad r_{x|xu}(x_1|x_0,u_1), \qquad r_{u|x}(u_1|x_0), \tag{D.6a}$$

with normalization constraints

$$\int r_{y|x\theta}(y_1|x_1,\theta)\,\mathrm{d}y_1 = 1 \quad \forall (x_1,\theta), \tag{D.7a}$$

$$\int r_{y|x}(y_1|x_1)\,\mathrm{d}y_1 = 1 \quad \forall x_1, \tag{D.7b}$$

$$\int r_{x|xu}(x_1|x_0,u_1)\,\mathrm{d}x_1 = 1 \quad \forall (x_0,u_1), \tag{D.7c}$$

$$\int r_{u|x}(u_1|x_0)\,\mathrm{d}u_1 = 1 \quad \forall x_0. \tag{D.7d}$$

We also use the derived marginals

$$q_{\mathrm{sep}}(x_1,\theta) := \int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}y_1, \tag{D.8a}$$

$$q_{yx}(y_1,x_1) := \int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}\theta, \tag{D.8b}$$

$$q_{\mathrm{trip}}(x_1,x_0,u_1) := \int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}\theta, \tag{D.8c}$$

$$q_{\mathrm{pair}}(x_0,u_1) := \int q_{\mathrm{trip}}(x_1,x_0,u_1)\,\mathrm{d}x_1, \tag{D.8d}$$

$$q_{ux}(u_1,x_0) := \int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_1\,\mathrm{d}\theta. \tag{D.8e}$$

The last quantity is just $q_{ux}(u_1,x_0) = q_{\mathrm{pair}}(x_0,u_1)$, but it is convenient to name it separately when deriving the policy channel.

D.2 COMBINED OBJECTIVE AND MODIFIED KERNELS

For $T = 1$, the correction added to the Bethe free energy is

$$\Delta F_{\mathrm{comb}} = \underbrace{2\mathbb{H}\left[q(y_1|x_1,\theta)\right] - \mathbb{H}\left[q(x_1|x_0,u_1)\right] - \mathbb{H}\left[q(y_1|x_1)\right]}_{\Delta^{\mathrm{AIF}}} + \underbrace{\mathbb{H}\left[q(u_1|x_0)\right]}_{\Delta^{\mathrm{planning}}}. \tag{D.9}$$

Using the variational characterization of conditional entropy, each conditional entropy is reparameterized by a channel. The signs matter:

- $+\mathbb{H}\left[q(u_1|x_0)\right]$ contributes $-\mathbb{E}_{q_{\mathrm{dyn}}}\left[\log r_{u|x}(u_1|x_0)\right]$, so $r_{u|x}$ enters in the numerator of the dynamics kernel.
- $-\mathbb{H}\left[q(x_1|x_0,u_1)\right]$ contributes $+\mathbb{E}_{q_{\mathrm{dyn}}}\left[\log r_{x|xu}(x_1|x_0,u_1)\right]$, so $r_{x|xu}$ enters in the denominator of the dynamics kernel.
- $+2\mathbb{H}\left[q(y_1|x_1,\theta)\right]$ contributes $-2\mathbb{E}_{q_{\mathrm{obs}}}\left[\log r_{y|x\theta}(y_1|x_1,\theta)\right]$, so $r^2_{y|x\theta}$ enters in the numerator of the observation kernel.
- $-\mathbb{H}\left[q(y_1|x_1)\right]$ contributes $+\mathbb{E}_{q_{yx}}\left[\log r_{y|x}(y_1|x_1)\right]$, so $r_{y|x}$ enters in the denominator of the observation kernel.

The resulting objective is

$$\begin{aligned} F_{\mathrm{comb}}[q,r] = &\int q_{\mathrm{obs}} \log \frac{q_{\mathrm{obs}}}{p(y_1|x_1,\theta)} \,\mathrm{d}y_1 \,\mathrm{d}x_1 \,\mathrm{d}\theta - 2\int q_{\mathrm{obs}} \log r_{y|x\theta} \,\mathrm{d}y_1 \,\mathrm{d}x_1 \,\mathrm{d}\theta \\ &+ \int q_{yx}(y_1,x_1) \log r_{y|x}(y_1|x_1) \,\mathrm{d}y_1 \,\mathrm{d}x_1 \\ &+ \int q_{\mathrm{dyn}} \log \frac{q_{\mathrm{dyn}}}{p(x_1|x_0,\theta,u_1)} \,\mathrm{d}x_1 \,\mathrm{d}x_0 \,\mathrm{d}\theta \,\mathrm{d}u_1 \\ &+ \int q_{\mathrm{dyn}} \log r_{x|xu}(x_1|x_0,u_1) \,\mathrm{d}x_1 \,\mathrm{d}x_0 \,\mathrm{d}\theta \,\mathrm{d}u_1 \\ &- \int q_{\mathrm{dyn}} \log r_{u|x}(u_1|x_0) \,\mathrm{d}x_1 \,\mathrm{d}x_0 \,\mathrm{d}\theta \,\mathrm{d}u_1 \\ &- \int q_{x_0} \log p(x_0) \,\mathrm{d}x_0 - \int q_{u_1} \log p(u_1) \,\mathrm{d}u_1 - \int q_{y_1} \log \hat{p}_y(y_1) \,\mathrm{d}y_1 \\ &+ (d_\theta - 1)\mathbb{H}\left[q_\theta\right] - \int q_\theta \log p(\theta) \,\mathrm{d}\theta \\ &+ (d_{x_1} - 1)\mathbb{H}\left[q_{x_1}\right] - \int q_{x_1} \log \hat{p}_x(x_1) \,\mathrm{d}x_1. \end{aligned} \tag{D.10}$$

The variable degrees are

$$d_\theta = 3, \qquad d_{x_0} = 2, \qquad d_{x_1} = 3, \qquad d_{y_1} = 2, \qquad d_{u_1} = 2. \tag{D.11}$$

Remark 11 (Combined kernels)**.** *The two non-singleton factors are reweighted by*

$$\tilde{f}_{\mathrm{obs}}(y_1,x_1,\theta) = \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)}, \tag{D.12a}$$

$$\tilde{f}^{\mathrm{comb}}_{\mathrm{dyn}}(x_1,x_0,\theta,u_1) = \frac{p(x_1|x_0,\theta,u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0,u_1)}. \tag{D.12b}$$

The observation kernel is

$$\tilde{f}_{\mathrm{obs}}(y_1,x_1,\theta) = \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)}, \tag{D.13}$$

while the dynamics kernel is a genuine ratio in which $r_{u|x}$ sharpens action selection and $r_{x|xu}$ spreads mass over predicted futures.

D.3 LAGRANGIAN

We introduce Lagrange multipliers for all normalization and marginalization constraints. The factor and consistency multipliers are

$$\lambda_{\mathrm{obs}}, \quad \lambda_{\mathrm{dyn}}, \quad \lambda_{y_1}(y_1), \quad \lambda^{(\mathrm{obs})}_{\theta}(\theta), \quad \lambda^{(\mathrm{obs})}_{x_1}(x_1), \quad \lambda^{(\mathrm{dyn})}_{x_1}(x_1), \quad \lambda_{x_0}(x_0), \quad \lambda_{u_1}(u_1), \quad \lambda^{(\mathrm{dyn})}_{\theta}(\theta), \tag{D.14}$$

and the channel multipliers are

$$\nu_{\mathrm{obs}}(x_1,\theta), \qquad \nu_{y|x}(x_1), \qquad \nu_x(x_0,u_1), \qquad \nu_{u|x}(x_0). \tag{D.15}$$

The full Lagrangian is

$$\begin{aligned}
\mathcal{L}_{\mathrm{comb}} = F_{\mathrm{comb}}[q,r] \\
&+ \lambda_{\mathrm{obs}} \left(\int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}y_1\,\mathrm{d}x_1\,\mathrm{d}\theta - 1 \right) \\
&+ \lambda_{\mathrm{dyn}} \left(\int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta\,\mathrm{d}u_1 - 1 \right) \\
&+ \int \lambda_{y_1}(y_1) \left(\int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}x_1\,\mathrm{d}\theta - q_{y_1}(y_1) \right) \mathrm{d}y_1 \\
&+ \int \lambda_{\theta}^{(\mathrm{obs})}(\theta) \left(\int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}y_1\,\mathrm{d}x_1 - q_{\theta}(\theta) \right) \mathrm{d}\theta \\
&+ \int \lambda_{x_1}^{(\mathrm{obs})}(x_1) \left(\int q_{\mathrm{obs}}(y_1,x_1,\theta)\,\mathrm{d}y_1\,\mathrm{d}\theta - q_{x_1}(x_1) \right) \mathrm{d}x_1 \\
&+ \int \lambda_{x_1}^{(\mathrm{dyn})}(x_1) \left(\int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_0\,\mathrm{d}\theta\,\mathrm{d}u_1 - q_{x_1}(x_1) \right) \mathrm{d}x_1 \\
&+ \int \lambda_{x_0}(x_0) \left(\int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_1\,\mathrm{d}\theta\,\mathrm{d}u_1 - q_{x_0}(x_0) \right) \mathrm{d}x_0 \\
&+ \int \lambda_{u_1}(u_1) \left(\int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta - q_{u_1}(u_1) \right) \mathrm{d}u_1 \\
&+ \int \lambda_{\theta}^{(\mathrm{dyn})}(\theta) \left(\int q_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}u_1 - q_{\theta}(\theta) \right) \mathrm{d}\theta \\
&+ \iint \nu_{\mathrm{obs}}(x_1,\theta) \left(\int r_{y|x\theta}(y_1|x_1,\theta)\,\mathrm{d}y_1 - 1 \right) \mathrm{d}x_1\,\mathrm{d}\theta \\
&+ \iint \nu_{y|x}(x_1) \left(\int r_{y|x}(y_1|x_1)\,\mathrm{d}y_1 - 1 \right) \mathrm{d}x_1 \\
&+ \iint \nu_{x}(x_0,u_1) \left(\int r_{x|xu}(x_1|x_0,u_1)\,\mathrm{d}x_1 - 1 \right) \mathrm{d}x_0\,\mathrm{d}u_1 \\
&+ \int \nu_{u|x}(x_0) \left(\int r_{u|x}(u_1|x_0)\,\mathrm{d}u_1 - 1 \right) \mathrm{d}x_0.
\end{aligned} \tag{D.16}$$

D.4 STATIONARITY CONDITIONS

D.4.1 Variation with respect to q_{obs}

The stationarity equation for the observation factor is

Proposition 12 (Observation factor belief for the combined objective)**.** *At stationarity,*

$$q^*_{\mathrm{obs}}(y_1,x_1,\theta) \propto \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)} e^{-\lambda_{y_1}(y_1)} e^{-\lambda_{\theta}^{(\mathrm{obs})}(\theta)} e^{-\lambda_{x_1}^{(\mathrm{obs})}(x_1)}. \tag{D.17}$$

D.4.2 Variation with respect to q_{dyn}

The dynamics-side terms now contain both channel contributions:

$$\begin{aligned}
&\int q_{\mathrm{dyn}} \log q_{\mathrm{dyn}} - \int q_{\mathrm{dyn}} \log p(x_1|x_0,\theta,u_1) + \int q_{\mathrm{dyn}} \log r_{x|xu}(x_1|x_0,u_1) \\
&\quad - \int q_{\mathrm{dyn}} \log r_{u|x}(u_1|x_0) + (\text{multiplier terms}).
\end{aligned} \tag{D.18}$$

Taking $\frac{\delta \mathcal{L}_{\mathrm{comb}}}{\delta q_{\mathrm{dyn}}} = 0$ gives

$$\log q_{\mathrm{dyn}} + 1 - \log p(x_1|x_0,\theta,u_1) + \log r_{x|xu}(x_1|x_0,u_1) - \log r_{u|x}(u_1|x_0) + \lambda_{\mathrm{dyn}} + \lambda_{x_1}^{(\mathrm{dyn})} + \lambda_{x_0} + \lambda_{u_1} + \lambda_{\theta}^{(\mathrm{dyn})} = 0. \tag{D.19}$$

Proposition 13 (Combined dynamics factor belief)**.** *At stationarity,*

$$q^*_{\mathrm{dyn}}(x_1, x_0, \theta, u_1) \propto \frac{p(x_1|x_0, \theta, u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0, u_1)} e^{-\lambda^{(\mathrm{dyn})}_{x_1}(x_1)} e^{-\lambda_{x_0}(x_0)} e^{-\lambda_{u_1}(u_1)} e^{-\lambda^{(\mathrm{dyn})}_{\theta}(\theta)}. \tag{D.20}$$

The dynamics factor is therefore reweighted by a policy-dependent numerator together with a predictive denominator.

D.4.3 Variation with respect to the observation channels

The observation-side channels satisfy

$$r^*_{y|x\theta}(y_1|x_1, \theta) = \frac{q_{\mathrm{obs}}(y_1, x_1, \theta)}{q_{\mathrm{sep}}(x_1, \theta)} = q(y_1|x_1, \theta), \tag{D.21a}$$

$$r^*_{y|x}(y_1|x_1) = \frac{q_{yx}(y_1, x_1)}{q_{x_1}(x_1)} = q(y_1|x_1). \tag{D.21b}$$

D.4.4 Variation with respect to $r_{x|xu}$

The predictive dynamics channel is obtained from

$$\frac{\delta \mathcal{L}_{\mathrm{comb}}}{\delta r_{x|xu}(x_1|x_0, u_1)} = \frac{q_{\mathrm{trip}}(x_1, x_0, u_1)}{r_{x|xu}(x_1|x_0, u_1)} + \nu_x(x_0, u_1) = 0. \tag{D.22}$$

Proposition 14 (Predictive dynamics channel)**.** *At stationarity,*

$$r^*_{x|xu}(x_1|x_0, u_1) = \frac{q_{\mathrm{trip}}(x_1, x_0, u_1)}{q_{\mathrm{pair}}(x_0, u_1)} = q(x_1|x_0, u_1). \tag{D.23}$$

D.4.5 Variation with respect to $r_{u|x}$

Since $r_{u|x}$ does not depend on x_1 or θ, only the marginal $q_{ux}(u_1, x_0)$ matters:

$$-\int q_{ux}(u_1, x_0) \log r_{u|x}(u_1|x_0)\, \mathrm{d}u_1\, \mathrm{d}x_0 + \int \nu_{u|x}(x_0) \left(\int r_{u|x}(u_1|x_0)\, \mathrm{d}u_1 - 1 \right) \mathrm{d}x_0. \tag{D.24}$$

Taking a pointwise derivative yields

$$-\frac{q_{ux}(u_1, x_0)}{r_{u|x}(u_1|x_0)} + \nu_{u|x}(x_0) = 0. \tag{D.25}$$

Normalization implies $\nu_{u|x}(x_0) = q_{x_0}(x_0)$.

Proposition 15 (Policy channel)**.** *At stationarity,*

$$r^*_{u|x}(u_1|x_0) = \frac{q_{ux}(u_1, x_0)}{q_{x_0}(x_0)} = q(u_1|x_0). \tag{D.26}$$

D.4.6 Variation with respect to singleton beliefs

The singleton calculations are elementary because they depend only on degree counting. Since

$$d_{y_1} = d_{x_0} = d_{u_1} = 2, \qquad d_\theta = d_{x_1} = 3, \tag{D.27}$$

the degree-2 and degree-3 variables behave differently. In particular:

- for degree-2 variables y_1, x_0, and u_1, the KL term and the singleton entropy cancel, so

$$\lambda_{y_1}(y_1) = -\log \hat{p}_y(y_1), \qquad \lambda_{x_0}(x_0) = -\log p(x_0), \qquad \lambda_{u_1}(u_1) = -\log p(u_1); \tag{D.28}$$

- for degree-3 variables θ and x_1, one obtains only constraints on sums of multipliers,

$$\lambda^{(\mathrm{obs})}_\theta(\theta) + \lambda^{(\mathrm{dyn})}_\theta(\theta) = -\log q^*_\theta(\theta) - 1 - \log p(\theta), \tag{D.29}$$

$$\lambda^{(\mathrm{obs})}_{x_1}(x_1) + \lambda^{(\mathrm{dyn})}_{x_1}(x_1) = -\log q^*_{x_1}(x_1) - 1 - \log \hat{p}_x(x_1). \tag{D.30}$$

The extra policy channel does not alter these facts, because it modifies only the non-singleton dynamics factor and does not introduce any new local-consistency constraint involving singletons.

D.5 SOLVING THE STATIONARITY EQUATIONS AS MESSAGE PASSING

Once the kernels (D.12) are identified, the rest of the derivation follows the standard Bethe logic. For a factor a with kernel $\tilde{f}_a$ and neighboring variables $\mathcal{E}(a)$, the factor-to-variable update is

$$\mu_{a\to j}(s_j) \propto \int \tilde{f}_a(\boldsymbol{s}_a) \prod_{i\in\mathcal{E}(a)\setminus j} \mu_{i\to a}(s_i)\,\mathrm{d}\boldsymbol{s}_{a\setminus j}. \tag{D.31}$$

D.5.1 Observation-factor messages

Using the observation kernel in (D.12a), the observation-factor messages are

$$\mu_{\mathrm{obs}\to\theta}(\theta) = \iint \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)} \mu_{y_1\to\mathrm{obs}}(y_1)\mu_{x_1\to\mathrm{obs}}(x_1)\,\mathrm{d}y_1\,\mathrm{d}x_1, \tag{D.32a}$$

$$\mu_{\mathrm{obs}\to x_1}(x_1) = \iint \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)} \mu_{y_1\to\mathrm{obs}}(y_1)\mu_{\theta\to\mathrm{obs}}(\theta)\,\mathrm{d}y_1\,\mathrm{d}\theta, \tag{D.32b}$$

$$\mu_{\mathrm{obs}\to y_1}(y_1) = \iint \frac{p(y_1|x_1,\theta)\, r^2_{y|x\theta}(y_1|x_1,\theta)}{r_{y|x}(y_1|x_1)} \mu_{x_1\to\mathrm{obs}}(x_1)\mu_{\theta\to\mathrm{obs}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}\theta. \tag{D.32c}$$

D.5.2 Dynamics-factor messages

The combined dynamics kernel yields

$$\mu_{\mathrm{dyn}\to\theta}(\theta) = \iiint \frac{p(x_1|x_0,\theta,u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0,u_1)} \mu_{x_1\to\mathrm{dyn}}(x_1)\mu_{x_0\to\mathrm{dyn}}(x_0)\mu_{u_1\to\mathrm{dyn}}(u_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}u_1, \tag{D.33a}$$

$$\mu_{\mathrm{dyn}\to x_1}(x_1) = \iiint \frac{p(x_1|x_0,\theta,u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0,u_1)} \mu_{x_0\to\mathrm{dyn}}(x_0)\mu_{u_1\to\mathrm{dyn}}(u_1)\mu_{\theta\to\mathrm{dyn}}(\theta)\,\mathrm{d}x_0\,\mathrm{d}u_1\,\mathrm{d}\theta, \tag{D.33b}$$

$$\mu_{\mathrm{dyn}\to x_0}(x_0) = \iiint \frac{p(x_1|x_0,\theta,u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0,u_1)} \mu_{x_1\to\mathrm{dyn}}(x_1)\mu_{u_1\to\mathrm{dyn}}(u_1)\mu_{\theta\to\mathrm{dyn}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}u_1\,\mathrm{d}\theta, \tag{D.33c}$$

$$\mu_{\mathrm{dyn}\to u_1}(u_1) = \iiint \frac{p(x_1|x_0,\theta,u_1)\, r_{u|x}(u_1|x_0)}{r_{x|xu}(x_1|x_0,u_1)} \mu_{x_1\to\mathrm{dyn}}(x_1)\mu_{x_0\to\mathrm{dyn}}(x_0)\mu_{\theta\to\mathrm{dyn}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta. \tag{D.33d}$$

D.5.3 Factor beliefs and singleton beliefs

The factor beliefs are kernel times incoming messages:

$$q^*_{\mathrm{obs}}(y_1,x_1,\theta) \propto \tilde{f}_{\mathrm{obs}}(y_1,x_1,\theta)\,\mu_{y_1\to\mathrm{obs}}(y_1)\mu_{x_1\to\mathrm{obs}}(x_1)\mu_{\theta\to\mathrm{obs}}(\theta), \tag{D.34a}$$

$$q^*_{\mathrm{dyn}}(x_1,x_0,\theta,u_1) \propto \tilde{f}^{\mathrm{comb}}_{\mathrm{dyn}}(x_1,x_0,\theta,u_1)\,\mu_{x_1\to\mathrm{dyn}}(x_1)\mu_{x_0\to\mathrm{dyn}}(x_0)\mu_{\theta\to\mathrm{dyn}}(\theta)\mu_{u_1\to\mathrm{dyn}}(u_1). \tag{D.34b}$$

Singleton beliefs keep the usual sum-product form:

$$q^*_\theta(\theta) \propto p(\theta)\,\mu_{\mathrm{obs}\to\theta}(\theta)\,\mu_{\mathrm{dyn}\to\theta}(\theta), \tag{D.35a}$$

$$q^*_{x_1}(x_1) \propto \hat{p}_x(x_1)\,\mu_{\mathrm{obs}\to x_1}(x_1)\,\mu_{\mathrm{dyn}\to x_1}(x_1), \tag{D.35b}$$

$$q^*_{x_0}(x_0) \propto p(x_0)\,\mu_{\mathrm{dyn}\to x_0}(x_0), \tag{D.35c}$$

$$q^*_{u_1}(u_1) \propto p(u_1)\,\mu_{\mathrm{dyn}\to u_1}(u_1), \tag{D.35d}$$

$$q^*_{y_1}(y_1) \propto \hat{p}_y(y_1)\,\mu_{\mathrm{obs}\to y_1}(y_1). \tag{D.35e}$$

D.5.4 Channel updates

At a fixed point, all four channels recover the corresponding conditionals under the current factor beliefs:

$$r^*_{y|x\theta}(y_1|x_1,\theta) = q(y_1|x_1,\theta), \tag{D.36a}$$
$$r^*_{y|x}(y_1|x_1) = q(y_1|x_1), \tag{D.36b}$$
$$r^*_{x|xu}(x_1|x_0,u_1) = q(x_1|x_0,u_1), \tag{D.36c}$$
$$r^*_{u|x}(u_1|x_0) = q(u_1|x_0). \tag{D.36d}$$

Remark 16 (Dynamics-side min-max structure)**.** *The two dynamics channels act on different conditionals and with opposite signs. The denominator $r_{x|xu}$ maximizes $\mathbb{H}\left[q(x_1|x_0,u_1)\right]$, while the numerator $r_{u|x}$ minimizes $\mathbb{H}\left[q(u_1|x_0)\right]$. The combined dynamics factor therefore interpolates between commitment over actions and spreading over future states.*

D.6 GENERIC SCHEME FOR ARBITRARY T

The passage from $T = 1$ to arbitrary horizons is immediate because all entropy corrections are additive across time, so each time step receives its own local channels

$$r_{y|x\theta,t}(y_t|x_t,\theta), \quad r_{y|x,t}(y_t|x_t), \quad r_{x|xu,t}(x_t|x_{t-1},u_t), \quad r_{u|x,t}(u_t|x_{t-1}). \tag{D.37}$$

The per-time-step kernels are

$$\tilde{f}_{\mathrm{obs}_t}(y_t,x_t,\theta) = \frac{p(y_t|x_t,\theta)\, r^2_{y|x\theta,t}(y_t|x_t,\theta)}{r_{y|x,t}(y_t|x_t)}, \tag{D.38a}$$
$$\tilde{f}^{\mathrm{comb}}_{\mathrm{dyn}_t}(x_t,x_{t-1},\theta,u_t) = \frac{p(x_t|x_{t-1},\theta,u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1},u_t)}. \tag{D.38b}$$

The full multi-step scheme is obtained by applying the usual sum-product updates with the kernels in (D.38).

D.6.1 Messages from Observation Factor f_{obs_t}

$$\mu_{\mathrm{obs}_t\to\theta}(\theta) = \iint \frac{p(y_t|x_t,\theta)\, r^2_{y|x\theta,t}(y_t|x_t,\theta)}{r_{y|x,t}(y_t|x_t)} \mu_{y_t\to\mathrm{obs}_t}(y_t)\mu_{x_t\to\mathrm{obs}_t}(x_t)\,\mathrm{d}y_t\,\mathrm{d}x_t, \tag{D.39a}$$
$$\mu_{\mathrm{obs}_t\to x_t}(x_t) = \iint \frac{p(y_t|x_t,\theta)\, r^2_{y|x\theta,t}(y_t|x_t,\theta)}{r_{y|x,t}(y_t|x_t)} \mu_{y_t\to\mathrm{obs}_t}(y_t)\mu_{\theta\to\mathrm{obs}_t}(\theta)\,\mathrm{d}y_t\,\mathrm{d}\theta, \tag{D.39b}$$
$$\mu_{\mathrm{obs}_t\to y_t}(y_t) = \iint \frac{p(y_t|x_t,\theta)\, r^2_{y|x\theta,t}(y_t|x_t,\theta)}{r_{y|x,t}(y_t|x_t)} \mu_{x_t\to\mathrm{obs}_t}(x_t)\mu_{\theta\to\mathrm{obs}_t}(\theta)\,\mathrm{d}x_t\,\mathrm{d}\theta. \tag{D.39c}$$

D.6.2 Messages from Dynamics Factor f_{dyn_t}

$$\mu_{\mathrm{dyn}_t\to\theta}(\theta) = \iiint \frac{p(x_t|x_{t-1},\theta,u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1},u_t)} \times \mu_{x_t\to\mathrm{dyn}_t}(x_t)\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\mu_{u_t\to\mathrm{dyn}_t}(u_t)\,\mathrm{d}x_t\,\mathrm{d}x_{t-1}\,\mathrm{d}u_t, \tag{D.40a}$$

$$\mu_{\mathrm{dyn}_t\to x_t}(x_t) = \iiint \frac{p(x_t|x_{t-1},\theta,u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1},u_t)} \mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\mu_{u_t\to\mathrm{dyn}_t}(u_t)\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mathrm{d}x_{t-1}\,\mathrm{d}u_t\,\mathrm{d}\theta, \tag{D.40b}$$

$$\mu_{\mathrm{dyn}_t\to x_{t-1}}(x_{t-1}) = \iiint \frac{p(x_t|x_{t-1},\theta,u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1},u_t)} \mu_{x_t\to\mathrm{dyn}_t}(x_t)\mu_{u_t\to\mathrm{dyn}_t}(u_t)\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mathrm{d}x_t\,\mathrm{d}u_t\,\mathrm{d}\theta, \tag{D.40c}$$

$$\mu_{\mathrm{dyn}_t\to u_t}(u_t) = \iiint \frac{p(x_t|x_{t-1},\theta,u_t)\, r_{u|x,t}(u_t|x_{t-1})}{r_{x|xu,t}(x_t|x_{t-1},u_t)} \mu_{x_t\to\mathrm{dyn}_t}(x_t)\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mathrm{d}x_t\,\mathrm{d}x_{t-1}\,\mathrm{d}\theta. \tag{D.40d}$$

D.6.3 Factor Beliefs

$$q^*_{\mathrm{obs},t}(y_t,x_t,\theta) \propto \tilde{f}_{\mathrm{obs}_t}(y_t,x_t,\theta)\,\mu_{y_t\to\mathrm{obs}_t}(y_t)\mu_{x_t\to\mathrm{obs}_t}(x_t)\mu_{\theta\to\mathrm{obs}_t}(\theta), \tag{D.41a}$$

$$q^*_{\mathrm{dyn},t}(x_t,x_{t-1},\theta,u_t) \propto \tilde{f}^{\mathrm{comb}}_{\mathrm{dyn}_t}(x_t,x_{t-1},\theta,u_t)\,\mu_{x_t\to\mathrm{dyn}_t}(x_t)\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\mu_{\theta\to\mathrm{dyn}_t}(\theta)\mu_{u_t\to\mathrm{dyn}_t}(u_t). \tag{D.41b}$$

D.6.4 Channel Updates

At a fixed point, each time step has the four local channel updates

$$r^*_{y|x\theta,t}(y_t|x_t,\theta) = q_t(y_t|x_t,\theta), \tag{D.42a}$$

$$r^*_{y|x,t}(y_t|x_t) = q_t(y_t|x_t), \tag{D.42b}$$

$$r^*_{x|xu,t}(x_t|x_{t-1},u_t) = q_t(x_t|x_{t-1},u_t), \tag{D.42c}$$

$$r^*_{u|x,t}(u_t|x_{t-1}) = q_t(u_t|x_{t-1}). \tag{D.42d}$$

D.6.5 Singleton Beliefs

$$q^*_{x_t}(x_t) \propto \hat{p}_x(x_t)\,\mu_{\mathrm{obs}_t\to x_t}(x_t)\,\mu_{\mathrm{dyn}_t\to x_t}(x_t)\,\mu_{\mathrm{dyn}_{t+1}\to x_t}(x_t), \tag{D.43a}$$

$$q^*_\theta(\theta) \propto p(\theta)\prod_{\tau=1}^{T}\mu_{\mathrm{obs}_\tau\to\theta}(\theta)\,\mu_{\mathrm{dyn}_\tau\to\theta}(\theta), \tag{D.43b}$$

$$q^*_{u_t}(u_t) \propto p(u_t)\,\mu_{\mathrm{dyn}_t\to u_t}(u_t), \tag{D.43c}$$

$$q^*_{y_t}(y_t) \propto \hat{p}_y(y_t)\,\mu_{\mathrm{obs}_t\to y_t}(y_t). \tag{D.43d}$$

The boundary conditions follow the usual temporal-edge conventions: at $t=1$, $\mu_{x_0\to\mathrm{dyn}_1}(x_0)=p(x_0)$; at $t=T$, the message $\mu_{\mathrm{dyn}_{T+1}\to x_T}$ is absent.

D.7 INTERPRETATION

At a fixed point,

$$\tilde{f}^{\mathrm{comb}}_{\mathrm{dyn}_t}(x_t,x_{t-1},\theta,u_t) = \frac{p(x_t|x_{t-1},\theta,u_t)\,q(u_t|x_{t-1})}{q(x_t|x_{t-1},u_t)}. \tag{D.44}$$

This exposes the objective transparently: the numerator implements policy commitment, while the denominator implements the risk-sensitive correction over predicted futures. The observation factor simultaneously balances the two observation-side entropy terms through $r_{y|x\theta}$ and $r_{y|x}$.

E MESSAGE PASSING DERIVATION FOR VBP

This appendix derives the message-passing equations for the Value Belief Propagation (VBP) scheme that implements cross-entropy planning (Section 4.1). The factor graph and Bethe approximation are identical to the combined-objective derivation (Appendix D); only the entropy correction differs. VBP requires a single channel reparameterization, making the derivation substantially simpler.

E.1 COORDINATE SYSTEM

We use the same generative model, factor graph, Bethe coordinates, and local-polytope constraints as in Appendix D: factor beliefs $q_{\mathrm{obs}}(y_1, x_1, \theta)$ and $q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)$, plus singleton beliefs $q_\theta, q_{x_0}, q_{x_1}, q_{y_1}, q_{u_1}$.

The single difference is the channel set. VBP introduces one channel:

$$r_{u|x}(u_1|x_0), \quad \text{subject to} \quad \int r_{u|x}(u_1|x_0)\,\mathrm{d}u_1 = 1 \quad \forall\, x_0. \tag{E.1}$$

This channel parameterizes the conditional entropy of actions given states via:

$$\mathbb{H}\left[q(u_1|x_0)\right] = \min_{r_{u|x}} \mathbb{E}_{q_{\mathrm{pair}}(x_0,u_1)}\left[-\log r_{u|x}(u_1|x_0)\right], \tag{E.2}$$

where the minimum is attained at $r_{u|x}(u_1|x_0) = q(u_1|x_0)$, and

$$q_{\mathrm{pair}}(x_0, u_1) := \iint q_{\mathrm{dyn}}(x_1, x_0, \theta, u_1)\,\mathrm{d}x_1\,\mathrm{d}\theta \tag{E.3}$$

is the marginal of the dynamics factor belief over (x_0, u_1).

E.2 OBJECTIVE FUNCTION

The VBP objective adds the cross-entropy planning correction (8) to the usual Bethe free energy:

$$\Delta F_{\mathrm{VBP}} = +\mathbb{H}\left[q(u_1|x_0)\right]. \tag{E.4}$$

After channel reparameterization via (E.2):

$$\begin{aligned} F_{\mathrm{VBP}}[q, r_{u|x}] = &\int q_{\mathrm{obs}} \log \frac{q_{\mathrm{obs}}}{p(y_1|x_1,\theta)}\,\mathrm{d}y_1\,\mathrm{d}x_1\,\mathrm{d}\theta \\ &+ \int q_{\mathrm{dyn}} \log \frac{q_{\mathrm{dyn}}}{p(x_1|x_0,\theta,u_1)}\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta\,\mathrm{d}u_1 \\ &- \int q_{\mathrm{dyn}} \log r_{u|x}(u_1|x_0)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta\,\mathrm{d}u_1 \\ &+ (\text{the usual singleton Bethe terms}). \end{aligned} \tag{E.5}$$

Remark 17. *Since $+\mathbb{H}\left[q(u_1|x_0)\right]$ carries a positive sign, the channel reparameterization contributes $-\mathbb{E}_q\left[\log r_{u|x}\right]$ to the objective. Consequently, $r_{u|x}$ appears in the* numerator *of the dynamics kernel (as a multiplicative factor), in contrast to the AIF dynamics channel $r_{x|xu}$ which appears in the denominator* (13b)*.*

E.3 STATIONARITY CONDITIONS

We form the Lagrangian with the same normalization and consistency multipliers as in Appendix D.3, except that the three observation and predictive-dynamics channel multipliers are replaced by a single multiplier $\nu_{u|x}(x_0)$ for the channel normalization constraint (E.1).

E.3.1 Variation with respect to q_{obs}

The observation factor receives no channel correction. The stationarity condition is identical to standard Bethe:

$$q^*_{\mathrm{obs}}(y_1, x_1, \theta) \propto p(y_1|x_1, \theta)\, e^{-\lambda_{y_1}(y_1)}\, e^{-\lambda_\theta^{(\mathrm{obs})}(\theta)}\, e^{-\lambda_{x_1}^{(\mathrm{obs})}(x_1)}. \tag{E.6}$$

The observation kernel is simply $p(y_1|x_1, \theta)$, as in standard belief propagation.

E.3.2 Variation with respect to q_{dyn}

The q_{dyn}-dependent terms include the channel correction $-\int q_{\mathrm{dyn}} \log r_{u|x}(u_1|x_0)$. Taking $\frac{\delta \mathcal{L}}{\delta q_{\mathrm{dyn}}} = 0$:

$$\log q_{\mathrm{dyn}} + 1 - \log p(x_1|x_0, \theta, u_1) - \log r_{u|x}(u_1|x_0) + \lambda_{\mathrm{dyn}} + \lambda_{x_1}^{(\mathrm{dyn})} + \lambda_{x_0} + \lambda_{u_1} + \lambda_\theta^{(\mathrm{dyn})} = 0. \tag{E.7}$$

Proposition 18 (VBP dynamics factor belief)**.** *At stationarity:*

$$q^*_{\mathrm{dyn}}(x_1, x_0, \theta, u_1) \propto p(x_1|x_0, \theta, u_1)\, r_{u|x}(u_1|x_0)\, e^{-\lambda_{x_1}^{(\mathrm{dyn})}(x_1)}\, e^{-\lambda_{x_0}(x_0)}\, e^{-\lambda_{u_1}(u_1)}\, e^{-\lambda_\theta^{(\mathrm{dyn})}(\theta)}. \tag{E.8}$$

The product $\tilde{f}_{\mathrm{dyn}}(x_1, x_0, \theta, u_1) := p(x_1|x_0, \theta, u_1)\, r_{u|x}(u_1|x_0)$ *is the VBP dynamics kernel.*

E.3.3 Variation with respect to $r_{u|x}$

The $r_{u|x}$-dependent terms in the Lagrangian are:

$$-\int q_{\mathrm{pair}}(x_0, u_1) \log r_{u|x}(u_1|x_0)\, \mathrm{d}u_1\, \mathrm{d}x_0 + \int \nu_{u|x}(x_0) \left(\int r_{u|x}(u_1|x_0)\, \mathrm{d}u_1 - 1 \right) \mathrm{d}x_0. \tag{E.9}$$

Pointwise derivative:

$$-\frac{q_{\mathrm{pair}}(x_0, u_1)}{r_{u|x}(u_1|x_0)} + \nu_{u|x}(x_0) = 0. \tag{E.10}$$

Imposing normalization $\int r_{u|x}\, \mathrm{d}u_1 = 1$ yields $\nu_{u|x}(x_0) = q_{x_0}(x_0)$, where $q_{x_0}(x_0) = \int q_{\mathrm{pair}}(x_0, u_1)\, \mathrm{d}u_1$.

Proposition 19 (Action channel)**.** *At stationarity:*

$$r^*_{u|x}(u_1|x_0) = \frac{q_{\mathrm{pair}}(x_0, u_1)}{q_{x_0}(x_0)} = q(u_1|x_0), \tag{E.11}$$

the conditional from the dynamics factor belief.

E.3.4 Singleton stationarity

The singleton stationarity conditions are identical to the corresponding degree-counting argument in Appendix D.4: degree-2 multipliers are identified algebraically, while degree-3 multipliers satisfy the same constraint equations on sums of multipliers.

E.4 MESSAGE-PASSING EQUATIONS

Using the observation kernel $\tilde{f}_{\mathrm{obs}} = p(y_1|x_1, \theta)$ and the VBP dynamics kernel $\tilde{f}_{\mathrm{dyn}} = p(x_1|x_0, \theta, u_1)\, r_{u|x}(u_1|x_0)$, the sum-product messages follow from the standard factor-to-variable update. Variable-to-factor messages $\mu_{i \to a}$ are the usual products of all incoming factor-to-variable messages except the one from the recipient factor.

E.4.1 Messages from Observation Factor

Standard sum-product (no channel modification):

$$\mu_{\text{obs}\to\theta}(\theta) = \iint p(y_1|x_1,\theta)\,\mu_{y_1\to\text{obs}}(y_1)\,\mu_{x_1\to\text{obs}}(x_1)\,\mathrm{d}y_1\,\mathrm{d}x_1, \tag{E.12a}$$

$$\mu_{\text{obs}\to x_1}(x_1) = \iint p(y_1|x_1,\theta)\,\mu_{y_1\to\text{obs}}(y_1)\,\mu_{\theta\to\text{obs}}(\theta)\,\mathrm{d}y_1\,\mathrm{d}\theta, \tag{E.12b}$$

$$\mu_{\text{obs}\to y_1}(y_1) = \iint p(y_1|x_1,\theta)\,\mu_{\theta\to\text{obs}}(\theta)\,\mu_{x_1\to\text{obs}}(x_1)\,\mathrm{d}x_1\,\mathrm{d}\theta. \tag{E.12c}$$

E.4.2 Messages from Dynamics Factor

Using kernel $p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)$:

$$\mu_{\text{dyn}\to\theta}(\theta) = \iiint p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)\,\mu_{x_0\to\text{dyn}}(x_0)\,\mu_{u_1\to\text{dyn}}(u_1)\,\mu_{x_1\to\text{dyn}}(x_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}u_1, \tag{E.13a}$$

$$\mu_{\text{dyn}\to x_1}(x_1) = \iiint p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)\,\mu_{x_0\to\text{dyn}}(x_0)\,\mu_{u_1\to\text{dyn}}(u_1)\,\mu_{\theta\to\text{dyn}}(\theta)\,\mathrm{d}x_0\,\mathrm{d}u_1\,\mathrm{d}\theta, \tag{E.13b}$$

$$\mu_{\text{dyn}\to x_0}(x_0) = \iiint p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)\,\mu_{x_1\to\text{dyn}}(x_1)\,\mu_{u_1\to\text{dyn}}(u_1)\,\mu_{\theta\to\text{dyn}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}u_1\,\mathrm{d}\theta, \tag{E.13c}$$

$$\mu_{\text{dyn}\to u_1}(u_1) = \iiint p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)\,\mu_{x_0\to\text{dyn}}(x_0)\,\mu_{\theta\to\text{dyn}}(\theta)\,\mu_{x_1\to\text{dyn}}(x_1)\,\mathrm{d}x_1\,\mathrm{d}x_0\,\mathrm{d}\theta. \tag{E.13d}$$

E.4.3 Factor Beliefs and Channel Update

The factor beliefs are the kernel times all incoming variable-to-factor messages:

$$q^*_{\text{obs}}(y_1,x_1,\theta) \propto p(y_1|x_1,\theta)\,\mu_{y_1\to\text{obs}}(y_1)\,\mu_{x_1\to\text{obs}}(x_1)\,\mu_{\theta\to\text{obs}}(\theta), \tag{E.14a}$$

$$q^*_{\text{dyn}}(x_1,x_0,\theta,u_1) \propto p(x_1|x_0,\theta,u_1)\,r_{u|x}(u_1|x_0)\,\mu_{x_1\to\text{dyn}}(x_1)\,\mu_{x_0\to\text{dyn}}(x_0)\,\mu_{\theta\to\text{dyn}}(\theta)\,\mu_{u_1\to\text{dyn}}(u_1). \tag{E.14b}$$

The channel update follows from Proposition 19:

$$r^*_{u|x}(u_1|x_0) = \frac{q_{\text{pair}}(x_0,u_1)}{q_{x_0}(x_0)} = q(u_1|x_0), \quad \text{where } q_{\text{pair}}(x_0,u_1) = \iint q^*_{\text{dyn}}\,\mathrm{d}x_1\,\mathrm{d}\theta. \tag{E.15}$$

E.4.4 Singleton Beliefs

Singleton beliefs follow the standard sum-product rule: the product of all incoming factor-to-variable messages with the prior (or goal prior):

$$q^*_\theta(\theta) \propto p(\theta)\,\mu_{\text{obs}\to\theta}(\theta)\,\mu_{\text{dyn}\to\theta}(\theta), \tag{E.16a}$$

$$q^*_{x_1}(x_1) \propto \hat{p}_x(x_1)\,\mu_{\text{obs}\to x_1}(x_1)\,\mu_{\text{dyn}\to x_1}(x_1), \tag{E.16b}$$

$$q^*_{x_0}(x_0) \propto p(x_0)\,\mu_{\text{dyn}\to x_0}(x_0), \tag{E.16c}$$

$$q^*_{u_1}(u_1) \propto p(u_1)\,\mu_{\text{dyn}\to u_1}(u_1), \tag{E.16d}$$

$$q^*_{y_1}(y_1) \propto \hat{p}_y(y_1)\,\mu_{\text{obs}\to y_1}(y_1). \tag{E.16e}$$

E.5 GENERIC SCHEME FOR ARBITRARY T

The VBP scheme generalizes to arbitrary time horizons by introducing a time-local channel $r_{u|x,t}(u_t|x_{t-1})$ at each time step $t = 1,\dots,T$. As in Appendix D.6, the per-timestep additivity of the Lagrangian and the entropy correction $+\sum_t \mathbb{H}\left[q(u_t|x_{t-1})\right]$ yields the multi-step scheme directly from the $T{=}1$ derivation.

E.5.1 Messages from Observation Factor f_{obs_t}

Standard sum-product (no channel modification):

$$\mu_{\mathrm{obs}_t\to\theta}(\theta) = \iint p(y_t|x_t,\theta)\,\mu_{y_t\to\mathrm{obs}_t}(y_t)\,\mu_{x_t\to\mathrm{obs}_t}(x_t)\,\mathrm{d}y_t\,\mathrm{d}x_t, \tag{E.17a}$$

$$\mu_{\mathrm{obs}_t\to x_t}(x_t) = \iint p(y_t|x_t,\theta)\,\mu_{y_t\to\mathrm{obs}_t}(y_t)\,\mu_{\theta\to\mathrm{obs}_t}(\theta)\,\mathrm{d}y_t\,\mathrm{d}\theta, \tag{E.17b}$$

$$\mu_{\mathrm{obs}_t\to y_t}(y_t) = \iint p(y_t|x_t,\theta)\,\mu_{\theta\to\mathrm{obs}_t}(\theta)\,\mu_{x_t\to\mathrm{obs}_t}(x_t)\,\mathrm{d}x_t\,\mathrm{d}\theta. \tag{E.17c}$$

E.5.2 Messages from Dynamics Factor f_{dyn_t}

Using kernel $p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})$:

$$\begin{aligned}\mu_{\mathrm{dyn}_t\to\theta}(\theta) = \iiint\ & p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})\,\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1}) \\ & \times \mu_{u_t\to\mathrm{dyn}_t}(u_t)\,\mu_{x_t\to\mathrm{dyn}_t}(x_t)\,\mathrm{d}x_t\,\mathrm{d}x_{t-1}\,\mathrm{d}u_t,\end{aligned} \tag{E.18a}$$

$$\mu_{\mathrm{dyn}_t\to x_t}(x_t) = \iiint p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})\,\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\,\mu_{u_t\to\mathrm{dyn}_t}(u_t)\,\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mathrm{d}x_{t-1}\,\mathrm{d}u_t\,\mathrm{d}\theta, \tag{E.18b}$$

$$\mu_{\mathrm{dyn}_t\to x_{t-1}}(x_{t-1}) = \iiint p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})\,\mu_{x_t\to\mathrm{dyn}_t}(x_t)\,\mu_{u_t\to\mathrm{dyn}_t}(u_t)\,\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mathrm{d}x_t\,\mathrm{d}u_t\,\mathrm{d}\theta, \tag{E.18c}$$

$$\mu_{\mathrm{dyn}_t\to u_t}(u_t) = \iiint p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})\,\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\,\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mu_{x_t\to\mathrm{dyn}_t}(x_t)\,\mathrm{d}x_t\,\mathrm{d}x_{t-1}\,\mathrm{d}\theta. \tag{E.18d}$$

E.5.3 Factor Beliefs

$$q^*_{\mathrm{obs},t}(y_t,x_t,\theta) \propto p(y_t|x_t,\theta)\,\mu_{y_t\to\mathrm{obs}_t}(y_t)\,\mu_{x_t\to\mathrm{obs}_t}(x_t)\,\mu_{\theta\to\mathrm{obs}_t}(\theta), \tag{E.19a}$$

$$q^*_{\mathrm{dyn},t}(x_t,x_{t-1},\theta,u_t) \propto p(x_t|x_{t-1},\theta,u_t)\,r_{u|x,t}(u_t|x_{t-1})\,\mu_{x_t\to\mathrm{dyn}_t}(x_t)\,\mu_{x_{t-1}\to\mathrm{dyn}_t}(x_{t-1})\,\mu_{\theta\to\mathrm{dyn}_t}(\theta)\,\mu_{u_t\to\mathrm{dyn}_t}(u_t). \tag{E.19b}$$

E.5.4 Channel Updates

Each time step has its own channel update:

$$r^*_{u|x,t}(u_t|x_{t-1}) = q_t(u_t|x_{t-1}), \quad \text{where } q_t(u_t|x_{t-1}) = \frac{q_{\mathrm{pair},t}(x_{t-1},u_t)}{q_{x_{t-1}}(x_{t-1})}, \tag{E.20}$$

with $q_{\mathrm{pair},t}(x_{t-1},u_t) = \iint q^*_{\mathrm{dyn},t}\,\mathrm{d}x_t\,\mathrm{d}\theta$.

E.5.5 Singleton Beliefs

$$q^*_{x_t}(x_t) \propto \hat{p}_x(x_t)\,\mu_{\mathrm{obs}_t\to x_t}(x_t)\,\mu_{\mathrm{dyn}_t\to x_t}(x_t)\,\mu_{\mathrm{dyn}_{t+1}\to x_t}(x_t), \tag{E.21a}$$

$$q^*_\theta(\theta) \propto p(\theta)\prod_{\tau=1}^{T}\mu_{\mathrm{obs}_\tau\to\theta}(\theta)\,\mu_{\mathrm{dyn}_\tau\to\theta}(\theta), \tag{E.21b}$$

$$q^*_{u_t}(u_t) \propto p(u_t)\,\mu_{\mathrm{dyn}_t\to u_t}(u_t), \tag{E.21c}$$

$$q^*_{y_t}(y_t) \propto \hat{p}_y(y_t)\,\mu_{\mathrm{obs}_t\to y_t}(y_t). \tag{E.21d}$$

The boundary conditions follow the usual temporal-edge conventions: at $t = 1$, $\mu_{x_0 \to \mathrm{dyn}_1}(x_0) = p(x_0)$; at $t = T$, the message $\mu_{\mathrm{dyn}_{T+1} \to x_T}$ is absent.

E.6 PROPERTIES

No min-max structure. Unlike AIF, where opposing channels create a min-max optimization over beliefs, VBP has a single channel that enters only the dynamics kernel numerator. The joint optimization over $(q, r_{u|x})$ is a pure minimization problem, avoiding the convergence difficulties of AIF's saddle-point structure.

Convergence. The observation factor uses standard sum-product messages with no channel modification, so standard BP convergence guarantees apply to the observation side. The dynamics channel provides a single-sided correction. While damping may improve convergence in practice, the lack of opposing channels makes it less critical than for AIF.

Reduction. Setting $r_{u|x,t}$ to uniform for all t removes the entropy correction and recovers standard belief propagation (marginal inference). The VBP scheme thus continuously interpolates between marginal inference (uniform channel) and full cross-entropy planning (converged channel).

Fixed-point interpretation. At convergence, $r^*_{u|x,t}(u_t|x_{t-1}) = q(u_t|x_{t-1})$, so the VBP dynamics kernel becomes $p(x_t|x_{t-1}, \theta, u_t)\, q(u_t|x_{t-1})$. This reweights the state transition by the policy conditioned on the previous state, implementing the action commitment that cross-entropy planning encodes.

F EXPERIMENT DETAILS

This appendix provides full details for the experiments in Section 6.

F.1 FROZEN LAKE ENVIRONMENT

Grid layout. We adapt the classic Frozen Lake environment [Brockman et al., 2016, Towers et al., 2025] to include epistemic uncertainty by treating the hole layout as unknown. A $4{\times}4$ grid where the agent starts at the top-left cell and must reach the goal at the bottom-right cell. A subset of the remaining cells are holes; stepping into a hole terminates the episode with failure. The ice surface makes transitions stochastic: with probability $1 - p_{\text{slip}}$ the agent moves in the intended direction, and with probability $p_{\text{slip}}/3$ it slips to each of the three remaining directions, where $p_{\text{slip}} = 0.1$.

Configurations. The hole layout is the unknown parameter θ. We sample 15 hole configurations uniformly at random (with fixed seeds for reproducibility), each placing 2 holes on the grid (a fraction 0.2 of the 14 non-start/goal cells, truncated to an integer). The agent does not know the hole locations and must infer them from observations.

Observation model. The observation model has two modalities. First, $2n_{\text{pos}}$ position channels observe the agent's position and scan mode with near-deterministic precision $(0.999/0.001)$, so the agent always knows where it is. Second, n_{pos} grid-cell channels each provide a binary "hole/safe" reading for the corresponding cell. Unscanned grid-cell observations are corrupted by distance-dependent noise: noise $= \alpha_{\text{base}} + \alpha_{\text{range}} \cdot d/d_{\max}$, where d is the Manhattan distance from the agent to the cell. A low-noise reading directly constrains which configurations θ are consistent, making the observation model approximately unambiguous. A SCAN action switches all grid-cell observations to near-deterministic $(0.999/0.001)$ at the cost of one time step.

State, action, and observation spaces. States encode position and scan mode: $2 \times n_{\text{pos}}$ states total (e.g., 32 for a $4{\times}4$ grid). Actions are the four cardinal directions plus SCAN ($|\mathcal{U}| = 5$). Observations consist of $2n_{\text{pos}}$ position channels and n_{pos} grid-cell channels.

Priors. The goal prior $\hat{p}(x_T)$ is a softmax preference peaking at the bottom-right cell, with penalties for hole positions (varying per θ). The parameter prior is uniform over the 15 configurations: $p(\theta) = 1/15$. The initial state prior $p(x_0)$ places all mass on the top-left cell in unscanned mode. The action prior assigns weight 1 to each movement action and weight $c_{\text{scan}} = 0.1$ to SCAN, then normalizes: $p(u_t) = w_{u_t} / \sum_u w_u$, giving $p(\text{move}) \approx 0.244$ and $p(\text{SCAN}) \approx 0.024$.

Planning parameters. Planning horizon $T = 15$, fixed across all decision steps. All methods use 400 iterations. We run 1000 episodes per method with a maximum of 15 steps per episode. Episode i uses seed i for reproducibility.

F.2 ROCKSAMPLE ENVIRONMENT

Grid layout. We use the canonical RockSample $(4, 3)$ environment [Smith and Simmons, 2004] in the epistemic planning framework, treating rock quality as the unknown parameter θ. A 4×4 grid where the agent starts at the left edge and can exit via the right edge. Three rocks are placed at known grid positions; their quality (good or bad) is unknown.

Configurations. Rock quality defines the unknown parameter θ. With 3 rocks each having binary quality, there are $n_\theta = 8$ configurations. The agent does not know rock quality and must infer it from CHECK readings.

Observation model. The observation model has two components. First, position channels observe the agent's position with noise parameter $\alpha_{\text{pos}} = 0.1$. Second, rock-quality information is available only through CHECK actions: a CHECK of rock i returns a binary "good/bad" reading whose accuracy depends on the Euclidean distance d from the agent to that rock: $p(\text{correct} \mid d) = \frac{1}{2}\left(1 + 2^{-d/d_{1/2}}\right)$, where $d_{1/2} = 2$ is the half-efficiency distance. At $d = 0$ the reading is deterministic; at $d = d_{1/2}$ accuracy is 75%; as $d \to \infty$ the reading approaches chance. Outside CHECK actions the agent receives no rock information, so no action changes the observation model at future steps and novelty is the only source of epistemic value.

Actions and rewards. The agent has nine actions: four cardinal movements, one CHECK per rock, SAMPLE, and EXIT (move off the right edge). CHECK_i senses rock i at the cost of one time step, providing the explicit epistemic actions. SAMPLE collects the rock at the current position, yielding reward $+2$ (good rock) or penalty -3 (bad rock). EXIT gives a fixed reward $+1$. Movement is deterministic ($p_{\text{slip}} = 0$).

State, action, and observation spaces. States encode position: $n_{\text{pos}} = 16$ states for the 4×4 grid. Actions: four cardinal directions, three CHECK actions, SAMPLE, and EXIT ($|\mathcal{U}| = 9$). Observations consist of n_{pos} position channels and 3 binary rock-quality channels.

Priors. The goal prior $\hat{p}(x_T)$ is a softmax preference with temperature $\tau = 1$ peaking at EXIT cells, with penalties for remaining on the grid. The parameter prior is uniform over the 8 configurations: $p(\theta) = 1/8$. The action prior assigns weight 1 to each movement action and weight $c_{\text{exit}} = 0.5$ to EXIT, then normalizes.

Planning parameters. Planning horizon $T = 12$, fixed across all decision steps. All methods use 100 iterations. We run 1000 episodes per method with a maximum of 25 steps per episode. Episode i uses seed i for reproducibility.

Full results with confidence intervals. Table 3 reports RockSample results with 95% confidence intervals.

Table 3: RockSample results with 95% confidence intervals, averaged over 1000 episodes.

Method	Avg. reward	Retrieval (%)	Avg. steps
BP	1.00 [1.00, 1.00]	0.0	3.00
VBP	1.00 [1.00, 1.00]	0.0	3.00
RM-MP	1.00 [1.00, 1.00]	0.0	3.00
Nuijten-MP	1.00 [1.00, 1.00]	0.0	3.00
AIF-MP	4.01 [3.90, 4.12]	98.7	8.56

F.3 WUMPUS WORLD ENVIRONMENT

Grid layout. We adapt the classic Wumpus World environment [Russell and Norvig, 1995] to include epistemic uncertainty by treating the layout as unknown. We simplify the classic dynamics to isolate the epistemic challenge: the agent has no orientation or inventory and navigates by cardinal movement. Transitions slip to one of the three unintended directions with probability $p_{\text{slip}} = 0.01$, as in Frozen Lake. The agent starts at cell 0 and must reach the gold cell. The grid contains pits and a wumpus (both absorbing hazards that terminate the episode) and a single gold cell.

Configurations. The locations of pits, wumpus, and gold define the unknown parameter θ. We use 25 configurations sampled with fixed seeds. Each configuration places 4 pits, one wumpus, and one gold on the grid, excluding the agent's starting cell.

Observation model. The observation model has two components, both noisy when unscanned. Three binary feature channels detect adjacency to hazards: *breeze* fires if adjacent to a pit, *stench* fires if adjacent to the wumpus, and *glitter* fires if on the gold cell. Unscanned feature channels have true-positive probability $p_{\text{tp}} = 1 - \alpha_{\text{obs}}$ and false-positive probability $p_{\text{fp}} = 0.1\,\alpha_{\text{obs}}$. Additionally, n_{pos} position channels encode the agent's position, also noisy when unscanned. Observations are ambiguous: a breeze indicates a nearby pit but not which neighbor, and position uncertainty compounds this ambiguity. The agent must integrate evidence across multiple positions to triangulate hazard locations. A SCAN action switches the three feature channels to near-deterministic $(0.999/0.001)$ for a single step, at the cost of one time step; the scan mode decays after one step, and position channels are unaffected. Because the improved precision lasts only one step, scanning does not change the observation model at future steps: the value of a scan lies entirely in the information the resulting reading carries about θ.

State, action, and observation spaces. States encode position and a transient scan mode that decays after one step: $2 \times n_{\text{pos}}$ states total (e.g., 50 for a 5×5 grid). Actions are the four cardinal directions plus SCAN ($|\mathcal{U}| = 5$). Observations consist of 3 binary feature channels (breeze, stench, glitter) and n_{pos} position channels.

Priors. The goal prior $\hat{p}(x_T)$ is a softmax preference peaking at the gold cell for each θ, with penalties for pits and the wumpus. The parameter prior is uniform over the 25 configurations. The initial state prior $p(x_0)$ places all mass on position 0 in unscanned mode. The action prior assigns weight 1 to each movement action and weight $c_{\text{scan}} = 0.7$ to SCAN, then normalizes: $p(u_t) = w_{u_t} / \sum_u w_u$, giving $p(\text{move}) \approx 0.213$ and $p(\text{SCAN}) \approx 0.149$.

Planning parameters. Planning horizon $T = 8$, fixed across all decision steps. All methods use 150 iterations. SCAN costs one time step (same as Frozen Lake). We run 1000 episodes per method with a maximum of 16 steps per episode. Episode i uses seed i for reproducibility.

Representative trajectories. Figures 4 and 5 show two episodes that illustrate the behavioral signatures behind the aggregate Wumpus gap reported in Section 6.2. Each panel shows one method on a fixed layout, with the agent's path shaded by step order (light early, dark late) and SCAN actions marked at the cell where they were taken. The line below each panel reports the terminal reward, the number of steps, and the outcome.

BP and VBP never scan and stall near the start in both episodes. RM-MP moves more but without direction: a single early scan does not redirect it, and it times out mid-grid (Figure 4) or falls into a pit (Figure 5). Nuijten-MP scans but does not convert the evidence into safe progress: it advances only one cell before timing out (Figure 4) or walks into the pit beside the start (Figure 5). AIF-MP interleaves single scans with movement and is the only method that reaches the gold in either episode. Aggregate success rates nevertheless remain below half (Table 2); these episodes illustrate the qualitative separation, not uniform success.

F.4 COMMON IMPLEMENTATION DETAILS

Software framework. All tensor operations and message-passing routines use JAX [Bradbury et al., 2018] with JIT compilation. All planning and inference functions are decorated with `jax.jit`, with the planning horizon and number of iterations as compile-time constants.

Log-space computation. All internal messages, beliefs, and channels are stored as log-probabilities. A sentinel value of -10^{12} replaces $-\infty$ for numerical stability, and a safe logarithm floors its argument at 10^{-30} before taking the log. Conversion to probability space occurs only at final output via softmax.

Message damping. Channel-based methods (VBP, RM-MP, and AIF-MP) apply geometric damping (18), implemented as a linear interpolation in log-space followed by per-condition renormalization. Damping is applied to every channel the method uses (policy, dynamics, observation, and marginal observation) after each BP iteration. Structural zeros are preserved: a position remains at -10^{12} only if both old and new values are -10^{12}. For VBP, this reduces to damping the single policy channel; BP and Nuijten-MP do not use damping ($\lambda = 1.0$). The damping parameter λ is selected per method and environment from the convergence sweep described in Section F.5. Table 4 reports the selected values.

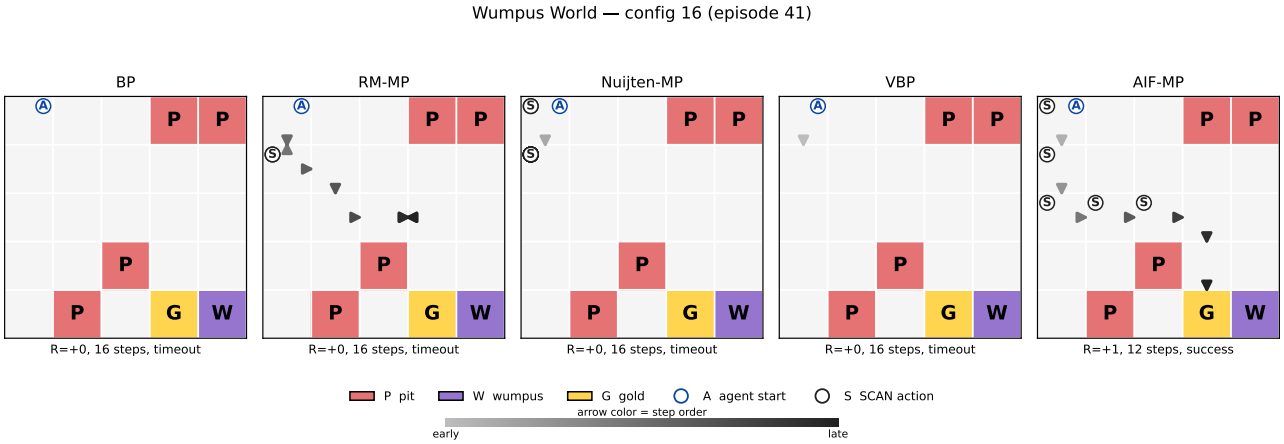


Figure 4: Wumpus World trajectories for all five methods on configuration 16, episode 41. Symbols: P pit, W wumpus, G gold, A agent start, S SCAN action; arrow shade encodes step order. BP and VBP never scan and stall near the start. RM-MP scans once but wanders to the middle of the grid and times out. Nuijten-MP scans twice yet advances only one cell. AIF-MP interleaves scans with movement and reaches the gold in 12 steps.

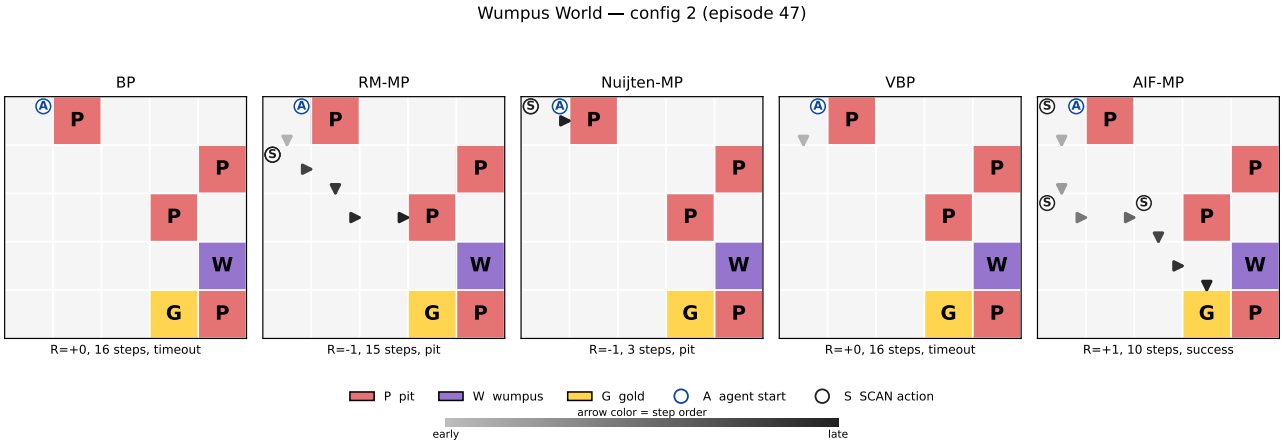


Figure 5: Wumpus World trajectories on configuration 2, episode 47, where the gold lies in the opposite corner of the grid. BP and VBP stall near the start. RM-MP wanders and falls into a pit after 15 steps. Nuijten-MP scans once and then walks into the pit beside the start. AIF-MP scans its way across the grid and reaches the gold in 10 steps.

Message initialization. All channels (dynamics, observation, and marginal observation) are initialized to uniform distributions over their respective domains. Parameter cavity beliefs are initialized to the prior $p(\theta)$; state beliefs are initialized to uniform over valid states.

Action selection. Actions are selected by greedy argmax over the action marginal at $t = 0$ (deterministic, no sampling).

F.5 CONVERGENCE BEHAVIOR

To select damping parameters and characterize convergence, we run a systematic sweep over $\lambda \in \{0.25, 0.4, 0.5, 0.6, 0.75, 0.9\}$ for each channel-based method (RM-MP, VBP, AIF-MP) and $\lambda = 1.0$ for loopy BP, across all three environments. Each configuration is run with 5 random seeds and 1,000 BP iterations. Convergence is declared when the maximum absolute change in any channel falls below 10^{-4}.

Convergence rate. Figure 6 reports the fraction of seeds that converge (color) and the median number of iterations to convergence (in parentheses) for each method–damping combination. The three environments exhibit qualitatively different

Table 4: Damping parameter λ per method and environment, selected from the convergence sweep (Section F.5).

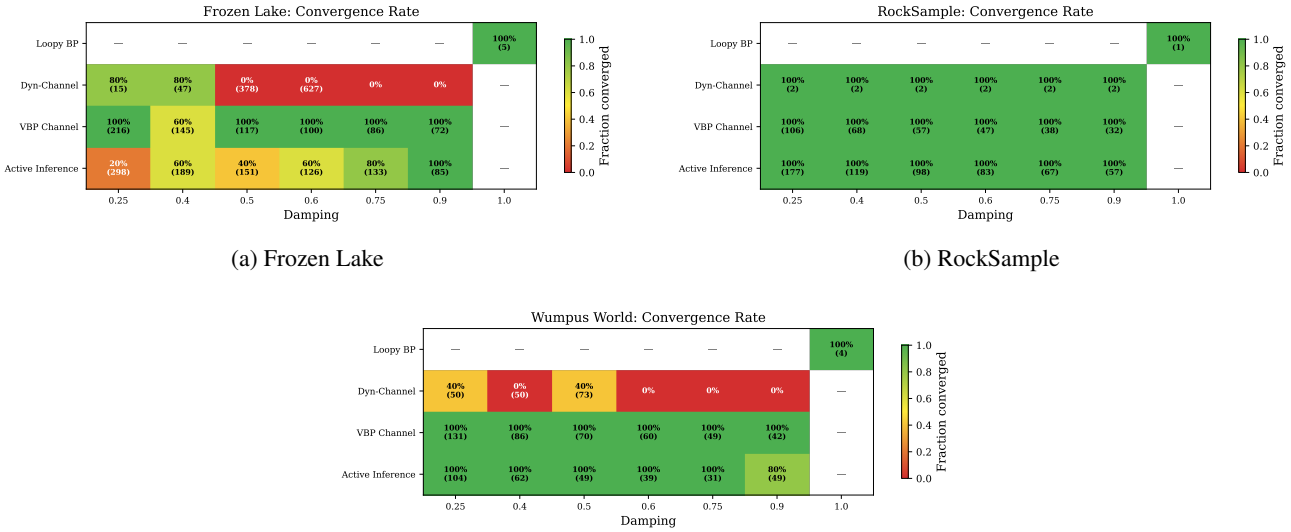
Method	Frozen Lake	RockSample	Wumpus World
BP	1.0	1.0	1.0
VBP	0.9	0.9	0.9
RM-MP	0.25	0.25	0.25
Nuijten-MP	1.0	1.0	1.0
AIF-MP	0.9	0.9	0.75

convergence profiles.

On RockSample (Figure 6b), all methods converge at all damping values. The deterministic dynamics make the dynamics channel update exact: it converges within two iterations regardless of λ, so the damping choice is immaterial for RM-MP here.

On Frozen Lake (Figure 6a), the dynamics channel (RM-MP) is the most fragile: it converges only at conservative damping (80% for $\lambda \leq 0.4$) and not at all for $\lambda \geq 0.5$, with large VFE oscillations. This is consistent with the min-max structure identified in Section 5: the stochastic dynamics activate the opposing-sign dynamics channel, which amplifies update steps when damping is insufficient. VBP is stable at nearly all damping values (60–100%), since it uses only the planning channel (no opposing signs). AIF-MP converges reliably at higher damping (100% at $\lambda = 0.9$) but slowly at conservative settings (20% at $\lambda = 0.25$).

On Wumpus World (Figure 6c), VBP and AIF-MP converge reliably: VBP at 100% for every damping value, AIF-MP at 100% for $\lambda \leq 0.75$ and 80% at $\lambda = 0.9$. The dynamics channel is again the most fragile, converging on at most 40% of seeds and on none for $\lambda \geq 0.6$. The one-step scan and the local adjacency signals keep the observation-side channels active, yet the joint scheme remains stable.



(a) Frozen Lake

(b) RockSample

(c) Wumpus World

Figure 6: Convergence rate (color) and median iterations to convergence (in parentheses) for each method and damping value λ, averaged over 5 seeds with $1{,}000$ iterations each. Dashes indicate that the method does not use damping at that value.

VFE convergence dynamics. Figure 7 shows VFE traces for all methods at their best damping on Frozen Lake. All four methods reach a stationary plateau within the iteration budget: loopy BP within $\sim$5 iterations, and the channel-augmented methods within 150 iterations, depending on the number of active channels. The absolute VFE values at the plateau are not directly comparable across methods because each method optimizes a different functional (Table 1).

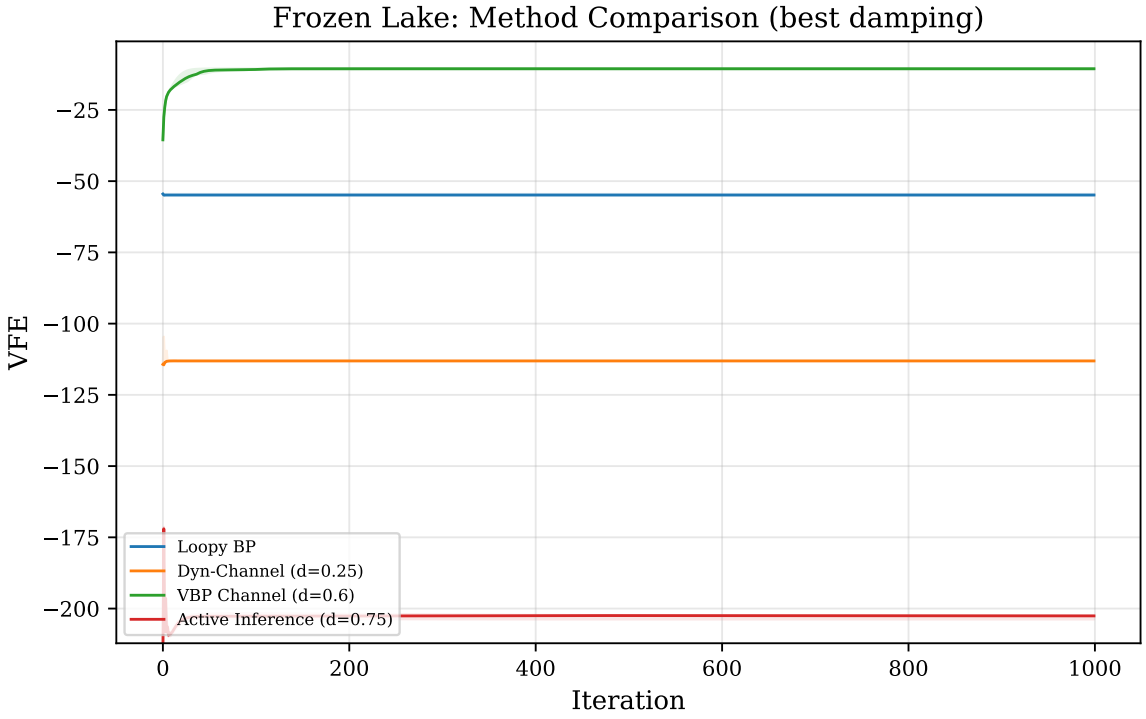


Figure 7: VFE traces on Frozen Lake for all methods at their best damping (seed-averaged with 1σ bands). All four methods reach a stationary plateau within the iteration budget.

Damping selection. The damping parameter λ is selected per method and environment as the value with the highest convergence rate; ties are broken by fewest median iterations. The selected values are reported in Table 4. The pattern is consistent across environments: RM-MP requires conservative damping ($\lambda = 0.25$) due to the opposing dynamics channel, VBP tolerates aggressive damping ($\lambda = 0.9$), and AIF-MP sits in between ($\lambda = 0.75$ on Wumpus World, 0.9 elsewhere). The worst observed case at the selected values is a single AIF-MP seed on Wumpus World that needs $\sim$320 iterations; all other runs converge within 150. Developing convergence theory for the channel-augmented scheme remains an open problem.