Formal Specification of a Discrete-Time Fine-Grained POMDP for Rodent Behavioral Analysis

Model Specification Document

November 25, 2025

Abstract

This document defines a Partially Observable Markov Decision Process (POMDP) designed to model rodent behavior in a probabilistic two-alternative forced choice task. Unlike standard trial-based models, this framework operates in discrete time ($\Delta t = 25$ms), capturing the micro-structure of behavior including lick trains, time-in-port, and task disengagement. The model is specifically constructed to infer three latent cognitive variables: **intrinsic exploration preference** (β), **information-seeking drive** (κ), and **decision noise** (τ).

1 Model Overview

The task is modeled as a discrete-time POMDP defined by the tuple $(\mathcal{C}, \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \gamma)$. The agent (rodent) maintains a belief distribution over hidden experimental contexts and selects actions to maximize a total value function that integrates extrinsic rewards, intrinsic exploration bonuses, and information-seeking incentives.

1.1 Time Discretization (Nyquist Compliance)

To guarantee the reconstruction of behavioral events with a minimum duration of 50ms (signal frequency ≈ 20Hz), the sampling frequency must satisfy the Nyquist rate (> 40Hz).

- **Time Step:** $\Delta t = 25$ms.
- **Implication:** A 50ms nose-poke is modeled as a sequence of exactly 2 discrete states ($s_{P1} \to s_{P2}$). This prevents aliasing errors where a 50ms event splits across bins.

2 The POMDP Formalism

2.1 Hidden Contexts ($\mathcal{C}$)

The experiment consists of distinct contexts, unobservable to the agent, which define the reward probabilities. To account for counterbalancing (where the high-reward side varies between animals), the context set is **symmetric**.

The set of contexts is $\mathcal{C} = \{C_T, C_{S2a}, C_{S2b}, C_{S3a}, C_{S3b}\}$.

2.2 Observable State Space ($\mathcal{S}$)

The state space captures the physical location and the sequential progress of the animal. The "In-Poke" state is split to enforce the duration requirement mechanistically. The "Armed" state

Context	Description	P(Reward | Spout 1)	P(Reward | Spout 2)
C_T	Training / Test 1	0.99	0.99
C_{S2a}	Test 2 (Right High)	0.50	0.99
C_{S2b}	Test 2 (Left High)	0.99	0.50
C_{S3a}	Test 3 (Right High)	0.25	0.50
C_{S3b}	Test 3 (Left High)	0.50	0.25

Table 1: Symmetric reward probabilities. *Note: Probabilities of 1.0 are replaced with 0.99 to prevent numerical singularities during Bayesian updates (the "Trembling Hand" assumption).*

is expanded to track cumulative licks on both spouts simultaneously.

$$\begin{aligned}
\mathcal{S} = \{s_I, &\quad \text{(Idle: Task available, not engaged)} \\
s_{P1}, &\quad \text{(Poke-Transient: First 25ms in port - Invalid for arming)} \\
s_{P2}, &\quad \text{(Poke-Valid: >25ms in port - Valid for arming)} \\
s_{k_1,k_2}, &\quad \text{(Armed \& Accumulating: } k_1 \text{ licks on S1, } k_2 \text{ licks on S2. } k_{1,2} \in \{0..4\}\text{)} \\
s_{C,R}, &\quad \text{(Consuming Reward: Event successful)} \\
s_{C,N} &\quad \text{(Consuming No-Reward: Event unsuccessful)}\}
\end{aligned}$$

Note: $s_{0,0}$ corresponds to the initial Armed state immediately after a valid poke.

2.3 Action Space ($\mathcal{A}$)

The set of micro-actions available at each time step.

- **Poke Actions**:
  - a_P (Engage Poke): Enter the port.
  - a_{SP} (Stay in Poke): Maintain head in port.
  - a_{LP} (Leave Poke): Exit the port.
- **Lick Actions**: a_{L1} (Lick Spout 1), a_{L2} (Lick Spout 2).
- **Wait Action**: a_W (Disengage/Wait). Used for exploration and pauses between licks.

2.4 Transition Dynamics ($\mathcal{T}$)

The transition function $\mathcal{T}(s'|s, a, c)$ defines the task mechanics.

2.4.1 Deterministic Mechanics (Context Independent)

Transitions involving poking and lick accumulation are deterministic.

1. Nose-Poke Logic (The 50ms Constraint) The duration requirement is enforced by the split state logic.

- **Initiation**: $\mathcal{T}(s_{P1}|s_I, a_P) = 1.0$.
- **Holding**: $\mathcal{T}(s_{P2}|s_{P1}, a_{SP}) = 1.0$ and $\mathcal{T}(s_{P2}|s_{P2}, a_{SP}) = 1.0$.
- **Abort**: $\mathcal{T}(s_I|s_{P1}, a_{LP}) = 1.0$. (Too short).
- **Success (Arming)**: $\mathcal{T}(s_{0,0}|s_{P2}, a_{LP}) = 1.0$. (Arms task, resets counters to 0).

2. Lick Logic (Parallel Accumulation) Licks accumulate on their respective counters without resetting the other.

- **Accumulate S1**: If $k_1 < 4$, increment k_1 and preserve k_2.

$$\mathcal{T}(s_{k_1+1,k_2} | s_{k_1,k_2}, a_{L1}) = 1.0$$

- **Accumulate S2**: If $k_2 < 4$, increment k_2 and preserve k_1.

$$\mathcal{T}(s_{k_1,k_2+1} | s_{k_1,k_2}, a_{L2}) = 1.0$$

- **Tolerance**: Waiting maintains the current counts (accommodating inter-lick intervals).

$$\mathcal{T}(s_{k_1,k_2} | s_{k_1,k_2}, a_W) = 1.0$$

- **Reset (Disengagement)**: Explicitly re-entering the poke or timing out (modeled as transition to Idle) resets the accumulation.

$$\mathcal{T}(s_{P1} | s_{k_1,k_2}, a_P) = 1.0$$

2.4.2 Stochastic Event Outcomes (Context Dependent)

The event triggers when **either** counter reaches 5 (transition from $k = 4$).

Trigger S1 (from s_{4,k_2}):

$$\mathcal{T}(s_{C,R} | s_{4,k_2}, a_{L1}, c) = P(\text{Reward} | \text{Spout 1}, c)$$
$$\mathcal{T}(s_{C,N} | s_{4,k_2}, a_{L1}, c) = 1 - P(\text{Reward} | \text{Spout 1}, c)$$

Trigger S2 (from $s_{k_1,4}$):

$$\mathcal{T}(s_{C,R} | s_{k_1,4}, a_{L2}, c) = P(\text{Reward} | \text{Spout 2}, c)$$
$$\mathcal{T}(s_{C,N} | s_{k_1,4}, a_{L2}, c) = 1 - P(\text{Reward} | \text{Spout 2}, c)$$

2.5 Reward Function ($\mathcal{R}$) and Intrinsic Motivation

The reward function is sparse to maintain interpretability. It includes both extrinsic (condensed milk) and intrinsic (exploration) components.

$$\mathcal{R}(s, a) = \begin{cases} 1 & \text{if } s = s_{C,R} \text{ (Extrinsic condensed milk reward)} \\ \beta & \text{if } s = s_I \text{ and } a = a_W \text{ (Intrinsic exploration reward)} \\ 0 & \text{otherwise} \end{cases}$$

Interpretation: β represents the value per time step (25ms) of strictly disengaging from the task.

2.6 Observations ($\Omega, \mathcal{O}$)

Since the outcome is encoded in the state ($s_{C,R}$ vs $s_{C,N}$), the observation function is an identity map on the states. The agent observes its state perfectly. The partial observability arises because the *transition probabilities* to the outcome states are unknown.

3 The Cognitive Agent

The agent is modeled as a Model-Based Reinforcement Learner that maintains a belief over contexts and computes values based on that belief.

3.1 Belief Initialization and Propagation

The model assumes the agent learns the task structure via experience across the longitudinal sequence of sessions.

1. **Naive Prior (Session 1):** At the onset of the very first training session, the agent holds a uniform belief over all contexts.
$$\mathbf{b}_{t=0}^{\text{sess}=1}(c) = \frac{1}{|\mathcal{C}|}$$
2. **Carry-Over (Subsequent Sessions):** To model the accumulated training history, the initial belief for session k is initialized as the final posterior belief of session $k-1$.
$$\mathbf{b}_{t=0}^{\text{sess}=k}(c) = \mathbf{b}_{T}^{\text{sess}=k-1}(c)$$

This mechanism allows the belief in the Training context (C_T) to naturally converge to near-certainty (≈ 1.0) prior to the first Testing session, without requiring artificial bias parameters.

3.2 Belief Updating

The agent maintains a belief vector $\mathbf{b}_t \in \Delta^{|\mathcal{C}|}$. Upon observing a transition to state s_{t+1} after taking action a_t, the belief is updated via Bayes' Rule:

$$\mathbf{b}_{t+1}(c) \propto \mathcal{T}(s_{t+1}|s_t, a_t, c) \cdot \mathbf{b}_t(c)$$

Note: For deterministic transitions (e.g., $s_{P1} \to s_{P2}$), the likelihood is 1.0 for all contexts, so the belief remains unchanged. Information is gained only at the stochastic outcome transitions.

3.3 Value Estimation (The Utility Approach)

The Q-value for an action a in belief state $\mathbf{b}_t$ is constructed from three components: expected extrinsic reward, intrinsic exploration bonus, and information-seeking bonus.

$$Q_{\text{total}}(\mathbf{b}_t, a) = Q_{\text{ext}}(\mathbf{b}_t, a) + \text{Bonus}_{\text{Explore}}(s, a) + \text{Bonus}_{\text{Info}}(\mathbf{b}_t, a) \tag{1}$$

3.3.1 1. Expected Extrinsic Value (Q_{ext})

The agent computes the weighted average of the optimal Q-values for each context (where $Q^*(c, a)$ is the value function of the underlying MDP for context c):

$$Q_{\text{ext}}(\mathbf{b}_t, a) = \sum_{c \in \mathcal{C}} \mathbf{b}_t(c) \cdot Q^*(c, a)$$

3.3.2 2. Intrinsic Exploration Bonus (β)

This term captures the drive to disengage from the task.

$$\text{Bonus}_{\text{Explore}}(s, a) = \begin{cases} \beta & \text{if } s = s_I, a = a_W \\ 0 & \text{otherwise} \end{cases}$$

3.3.3 3. Information-Seeking Bonus (κ)

This term captures directed exploration (curiosity). It is proportional to the expected reduction in belief entropy (Information Gain).

$$\text{Bonus}_{\text{Info}}(\mathbf{b}_t, a) = \kappa \cdot \mathbb{E}_{s'} \left[H(\mathbf{b}_t) - H(\mathbf{b}_{t+1}|s', a) \right]$$

where $H(\mathbf{b})$ is the Shannon entropy of the belief distribution.

3.4 Policy (Action Selection)

The agent selects actions stochastically using a Softmax decision rule:

$$P(a_t|\mathbf{b}_t) = \frac{\exp(Q_{\text{total}}(\mathbf{b}_t, a_t)/\tau)}{\sum_{a' \in \mathcal{A}} \exp(Q_{\text{total}}(\mathbf{b}_t, a')/\tau)}$$

Note: The temperature parameter τ determines the decision noise. $\tau \to 0$ implies deterministic greedy selection, while $\tau \to \infty$ implies uniform random selection.

4 Hierarchical Bayesian Fitting Procedure

To estimate the parameters $\theta = \{\beta, \kappa, \tau\}$ for Vehicle and Treatment groups, we employ a Hierarchical Bayesian Model (HBM).

4.1 Hierarchy Structure

1. **Group Level**: Hyperparameters define the population means for Vehicle and Treatment groups. Since temperature must be positive, it is modeled in log-space.

$$\begin{aligned} \mu_\beta^{\text{Veh}}, \mu_\beta^{\text{Treat}} &\sim \mathcal{N}(0, 1) \\ \mu_\kappa^{\text{Veh}}, \mu_\kappa^{\text{Treat}} &\sim \mathcal{N}(0, 1) \\ \mu_{\log\tau}^{\text{Veh}}, \mu_{\log\tau}^{\text{Treat}} &\sim \mathcal{N}(-1, 1) \quad \text{(Log-Normal prior)} \end{aligned}$$

2. **Subject Level**: Each animal i has individual parameters drawn from its group distribution.

$$\begin{aligned} \beta_i &\sim \mathcal{N}(\mu_\beta^{\text{Group}}, \sigma_\beta^{\text{Group}}) \\ \kappa_i &\sim \mathcal{N}(\mu_\kappa^{\text{Group}}, \sigma_\kappa^{\text{Group}}) \\ \log(\tau_i) &\sim \mathcal{N}(\mu_{\log\tau}^{\text{Group}}, \sigma_{\log\tau}^{\text{Group}}) \end{aligned}$$

3. **Subject Sequence Likelihood**: For a given subject i, observing a longitudinal sequence of M sessions (training followed by testing). For each session $j \in \{1..M\}$, actions $A^{(j)}$ and states $S^{(j)}$ are observed. The likelihood is the product over all steps in all sessions, respecting the belief propagation defined in Section 4.1:

$$\mathcal{L}(\{A, S\}|\beta_i, \kappa_i, \tau_i) = \prod_{j=1}^{M} \prod_{t=1}^{T_j} P_{\text{Softmax}}(a_t^{(j)}|\mathbf{b}_t^{(j)}(\beta_i, \kappa_i), \tau_i)$$

4.2 Inference

Posterior distributions are estimated using Hamiltonian Monte Carlo (HMC) via Stan/PyMC. The primary hypothesis tests compare the posterior distributions of the group means for all three cognitive components (Exploration, Information Seeking, and Decision Noise).