Reading Report:
A POMDP Formulation of Preference Elicitation Problems

[Your Name]

Paper Overview

Craig Boutilier (2002) casts the problem of *preference elicitation*—deciding which questions to ask a user so as to maximize decision quality while minimizing elicitation cost—as a *Partially Observable Markov Decision Process (POMDP).* Utility functions are treated as hidden states, queries as information-gathering actions (with cost), and final decisions as terminal actions (with reward equal to expected utility). This formulation subsumes earlier myopic value-of-information approaches by enabling multi-step lookahead, handling noisy responses, and supporting offline policy computation for arbitrary priors.

Key Contributions

1. **POMDP Modeling.** States are continuous beliefs over utility functions; actions comprise parameterized queries and terminal decisions; observations are noisy yes/no responses; and costs/rewards are formalized within the POMDP framework.
2. **Belief & Action Representation.** Proposes mixture models (Gaussian or uniform) for continuous belief states, and shows how queries update these into truncated mixtures.
3. **Approximate Solution Techniques.** Uses function approximation for query Q-functions $Q_i(l, \theta)$ (e.g. quadratic approximators), exploits the linearity of mixture Q-values, and employs gradient-based optimization over the query parameter l.
4. **Empirical Validation.** Demonstrates Bellman error of 2–3% on small decision tasks (4 outcomes, 6 decisions) and tractable offline computation for larger tasks (20 outcomes, 30 decisions), with fast sub-second online policy execution.

Reflections

1. **Generalization of Value-Function Approximation.** A quadratic function approximator achieves good performance in small problems, but when the number of outcomes or query types grows, is a quadratic model expressive enough? Would more flexible approximators (e.g. deep neural networks or Gaussian processes) yield better estimates of value of information?
2. **Adaptive Process Noise Estimation.** One key insight from the POMDP formulation is that, just as the agent must adaptively estimate its observation noise R based on the innovation in user responses, our Kalman-style optimizer should estimate its process noise covariance Q online to capture the true variability of loss and gradient dynamics. Concretely:

- **Innovation-based update:** Compute the innovation

  $$\delta_k = g_k - \hat{g}_k,$$

  where g_k is the observed gradient (or loss) at step k and $\hat{g}_k$ is its prediction from the previous state.
- **Exponential moving average:** Update Q via

  $$Q_{k+1} = \alpha\, Q_k + (1 - \alpha)\, \delta_k\, \delta_k^T,$$

  where $0 < \alpha < 1$ controls the memory of past innovations.
- **Responsive smoothing:** As training dynamics accelerate (e.g., during learning-rate warmup or near sharp minima), the innovation variance grows and Q increases, allowing the filter to be more responsive. Conversely, in flatter regions the innovation shrinks, reducing Q and enforcing smoother updates.