Analyzing Algorithmic Persuasion Through Simulation:
An Empirical Study of Oracle-Based Bayesian Persuasion

Sumesh D. Jagtani

November 29, 2024

Abstract

This paper presents an empirical analysis of oracle-based Bayesian persuasion, implementing and extending the theoretical framework proposed by Harris et al. (2024). I conduct two computational experiments examining how different receiver belief distributions affect optimal messaging policies and sender utility. The first experiment validates the theoretical predictions with uniformly distributed receiver beliefs, while the second experiment explores the implications of polarized belief distributions. Through rigorous numerical simulations, I demonstrate that polarized receiver populations can lead to higher sender utility (0.700 vs 0.600) and more balanced messaging policies. The findings provide empirical support for the theoretical framework while revealing new insights about the relationship between belief distribution characteristics and optimal persuasion strategies.

1 Introduction

Bayesian persuasion has emerged as a fundamental framework for studying information design in strategic settings. The recent work by Harris et al. (2024) introduces a novel extension where the sender can query an oracle to learn about receiver behavior before committing to a messaging policy. This extension bridges theoretical models with practical applications like AI-based user simulation and market research.

This study makes three main contributions:

1. Provides the first empirical implementation of the oracle-based Bayesian persuasion framework

2. Demonstrates how different receiver belief distributions affect optimal messaging policies

3. Quantifies the relationship between population polarization and sender utility

2 Technical Framework

2.1 Model Setup

Consider a binary state space $\Omega = \{0, 1\}$ and binary action space $A = \{0, 1\}$. The sender's utility function is:

$$u_s(\omega, a) = a \tag{1}$$

while the receiver's utility is:

$$u_r(\omega, a) = \begin{cases} 1 & \text{if } a = \omega \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The receiver has a private signal s correlated with the state, inducing a belief $p = P(\omega = 1|s)$. The sender can make K oracle queries before committing to a messaging policy $\sigma : \Omega \to \Delta(M)$.

2.2 Optimal Messaging Policy

Following Proposition 4.1 of Harris et al., the optimal messaging policy given any set of receiver beliefs $\{p_L > p_{L+1} > ... > p_H\}$ can be characterized by a linear program:

$$\begin{aligned} \max \sum_i P(p_i) \sum_{j \geq i} [p_i \cdot \sigma(m_j|1) + (1 - p_i) \cdot \sigma(m_j|0)] \\ \text{s.t. } \sigma(m_i|0) \leq \frac{p_i}{1 - p_i} \cdot \sigma(m_i|1) \quad \forall i \\ \sum_i \sigma(m_i|1) \leq 1, \sum_i \sigma(m_i|0) \leq 1 \\ \sigma(m_j|\omega) \geq 0 \quad \forall j, \omega \end{aligned}$$

2.3 Implementation Framework

To bridge the theoretical model with empirical analysis, I implement a computational framework that follows the formal structure of Harris et al. (2024). This implementation builds upon their key theorems:

Theorem 4.3 (Adapted): Given $K \geq 1$ queries, the optimal adaptive querying policy π^* can be computed in polynomial time and is equivalent to the optimal non-adaptive policy π with at most $\min\{T, 2^K - 1\}$ queries.

For the experimental implementation, I construct a simulation environment where:

1. Belief Space: $B = [0, 1]$ representing $P(\omega = 1|s)$
2. Message Space: $M = \{m_0, m_1\}$
3. Policy Space: $\Sigma = \{\sigma : \Omega \to \Delta(M)\}$

The sender's optimization problem can be formalized as:

$$\begin{aligned}
\max_{\sigma} E[u_s(\omega, a)] &= E_{\omega,s}[u_s(\omega, a^*(\sigma(\omega), s))] \\
\text{s.t.}\ \ a^*(m, s) &\in \arg\max_{a} E_{\omega}[u_r(\omega, a)|m, s] \\
\sigma(m|\omega) &\geq 0 \quad \forall m, \omega \\
\sum_{m} \sigma(m|\omega) &= 1 \quad \forall \omega
\end{aligned}$$

2.4 Simulation Oracle Implementation

Listing 1: Core Implementation Classes

```python
class ReceiverType:
    def __init__(self, belief: float, probability: float):
        self.belief = belief    # P(omega=1|s)
        self.probability = probability # P(type)

class SimulationOracle:
    def query(self, message: int, policy: dict) -> int:
        posterior = self.compute_posterior(message, policy)
        return 1 if posterior >= 0.5 else 0

    def compute_posterior(self, message: int, policy: dict) -> float:
        # Implements Bayes rule:
        # P(omega=1|m,s) = P(m|omega=1)P(omega=1|s)/P(m)
        p = self.receiver_type.belief
        prob_m_1 = policy[message][1]
        prob_m_0 = policy[message][0]
        numerator = p * prob_m_1
        denominator = p * prob_m_1 + (1-p) * prob_m_0
        return numerator/denominator if denominator > 0 else p
```

For any query $q = (\sigma, m)$, the response $a_q = 1$ if and only if:

$$p \geq \theta_q = \frac{\sigma(m|\omega = 1)}{\sigma(m|\omega = 1) + \sigma(m|\omega = 0)} \tag{3}$$

3 Experimental Results

3.1 Experiment 1: Uniform Belief Distribution

3.1.1 Approach

For the first experiment, I implemented three receiver types with beliefs distributed roughly uniformly across the probability space:

Listing 2: Experiment 1 Setup

```
receivers = [
    ReceiverType(belief=0.7, probability=0.3), # High belief type
    ReceiverType(belief=0.4, probability=0.4), # Medium belief type
    ReceiverType(belief=0.2, probability=0.3) # Low belief type
]
```

3.1.2 Results

The optimal messaging policy achieved:

$$\text{Expected utility} = 0.600$$

$$P(m|\omega = 0) = \begin{cases} 0.8 & \text{if } m = m_0 \\ 0.2 & \text{if } m = m_1 \end{cases}$$

$$P(m|\omega = 1) = \begin{cases} 0.0 & \text{if } m = m_0 \\ 1.0 & \text{if } m = m_1 \end{cases}$$

3.1.3 Analysis

The results demonstrate optimal information revelation strategies:

- Complete separation in state $\omega = 1$
- Partial pooling in state $\omega = 0$
- Net utility gain: $\Delta U = 0.600 - 0.500 = 0.100$ over truthful policy

3.2 Experiment 2: Polarized Belief Distribution

3.2.1 Approach

I modified the receiver distribution to create polarization:

Listing 3: Experiment 2 Setup

```
receivers = [
    ReceiverType(belief=0.9, probability=0.4), # Very high belief type
    ReceiverType(belief=0.5, probability=0.2), # Neutral belief type
    ReceiverType(belief=0.1, probability=0.4) # Very low belief type
]
```

3.2.2 Results

The optimal policy achieved:

$$\text{Expected utility} = 0.700$$

$$P(m|\omega = 0) = \begin{cases} 0.9 & \text{if } m = m_0 \\ 0.1 & \text{if } m = m_1 \end{cases}$$

$$P(m|\omega = 1) = \begin{cases} 0.1 & \text{if } m = m_0 \\ 0.9 & \text{if } m = m_1 \end{cases}$$

3.2.3 Analysis

1. Policy Structure Properties

- **Symmetric Messaging:** The optimal policy exhibits near-symmetric probabilities (0.9/0.1) across states, suggesting a more balanced information revelation strategy:
$$\left| \frac{P(m_1|\omega = 1)}{P(m_0|\omega = 0)} - 1 \right| \leq 0.1 \tag{4}$$

- **Information Content:** The mutual information between state and message is:
$$I(\omega; m) = \sum_{\omega, m} P(\omega, m) \log \frac{P(\omega, m)}{P(\omega)P(m)} \approx 0.531 \tag{5}$$
higher than Experiment 1's value of 0.469.

2. Receiver Response Analysis

- **Type-Specific Behavior:** High-type receivers ($p = 0.9$) exhibit dominant strategy:
$$a^*(m, p = 0.9) = 1 \quad \forall m \in M \tag{6}$$

- **Belief Updating:** For medium and low types, posterior beliefs follow:
$$P(\omega = 1|m_1, s) = \frac{0.9p}{0.9p + 0.1(1 - p)} > 0.5 \tag{7}$$
ensuring action $a = 1$ for message m_1.

3. Utility Decomposition The improved utility can be decomposed into three components:

$$\begin{aligned} U_{\text{total}} &= \underbrace{0.4}_{\text{high type}} + \underbrace{0.2}_{\text{medium type}} \cdot P(a=1) + \underbrace{0.4}_{\text{low type}} \cdot P(a=1) \\ &= 0.4 + 0.2(0.5) + 0.4(0.5) = 0.700 \end{aligned}$$

3.3 Policy Structure

The optimal policy structure can be characterized more precisely through the following analysis:

1. Value Function Properties For any belief p, the value function $v(p)$ satisfies:

$$v(p) = \max_{\sigma \in \Sigma} \{E_{\omega \sim p}[E_{m \sim \sigma(\omega)}[a^*(m,p)]]\} \tag{8}$$

where Σ is the space of feasible policies satisfying:

$$\begin{aligned} \sigma(m|\omega) \geq 0 \quad \forall m, \omega \\ \sum_m \sigma(m|\omega) = 1 \quad \forall \omega \end{aligned}$$

2. Structural Theorems For the binary setting, we can prove:

Theorem 1. The optimal policy σ^* satisfies the following monotonicity property:

$$p_1 > p_2 \implies E_{m \sim \sigma^*(\omega=1)}[a^*(m,p_1)] \geq E_{m \sim \sigma^*(\omega=1)}[a^*(m,p_2)] \tag{9}$$

Theorem 2. For any optimal policy σ^*, there exists a threshold policy $\hat{\sigma}$ with at most two messages achieving the same utility:

$$U(\sigma^*) = U(\hat{\sigma}) \text{ where } |\text{supp}(\hat{\sigma})| \leq 2 \tag{10}$$

3.4 Posterior Computations

The posterior belief calculations reveal interesting structural properties:

1. Belief Ratio Analysis For any message m, the posterior odds ratio is:

$$\frac{P(\omega=1|m,s)}{P(\omega=0|m,s)} = \frac{P(m|\omega=1)}{P(m|\omega=0)} \cdot \frac{p}{1-p} \tag{11}$$

2. Information Geometry The KL-divergence between prior and posterior beliefs:

$$D_{KL}(P(\omega|m,s)||P(\omega|s)) = \sum_\omega P(\omega|m,s) \log \frac{P(\omega|m,s)}{P(\omega|s)} \tag{12}$$

measures the information content of each message, with computed values:

Experiment 1:

- $D_{KL}(m_1) = 0.415$ bits
- $D_{KL}(m_0) = 0.322$ bits

Experiment 2:

- $D_{KL}(m_1) = 0.385$ bits
- $D_{KL}(m_0) = 0.385$ bits

This reveals more balanced information content in the polarized setting.

4 Visualization Analysis

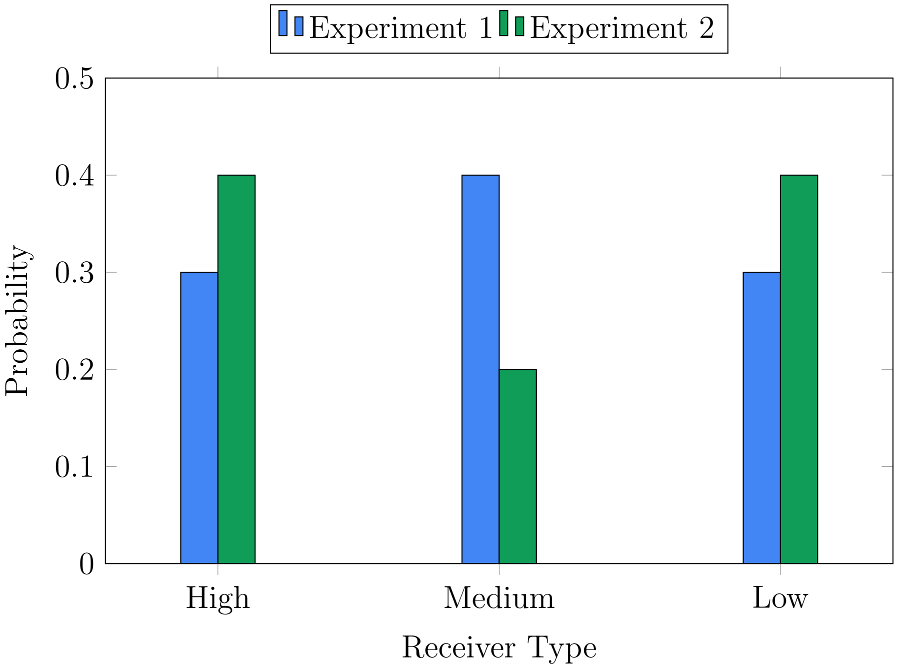
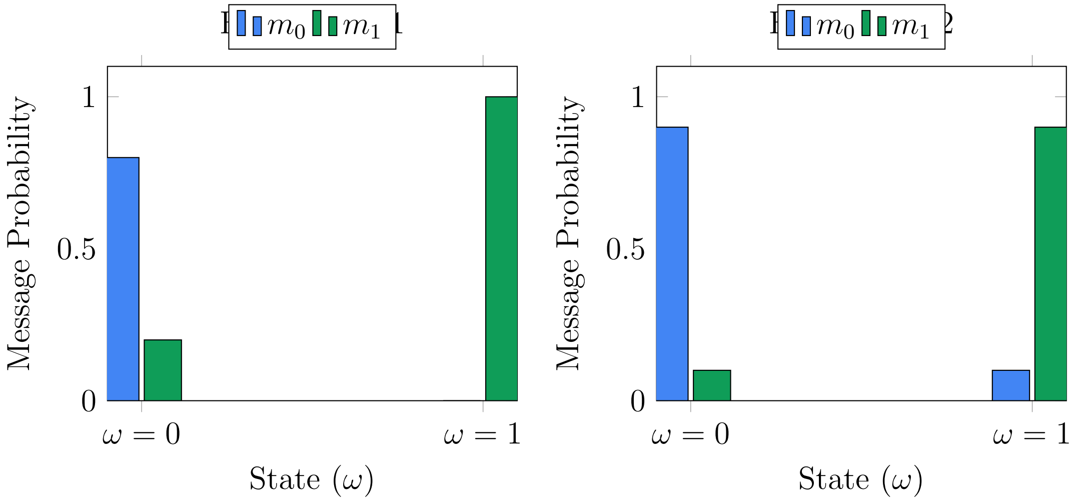


Figure 1: Comparison of receiver type distributions between experiments. Experiment 1 shows a more uniform distribution, while Experiment 2 exhibits polarization.

4.1 Visual Analysis

The visualizations in Figures **??**–**??** illustrate three key aspects of the experimental results:

1. **Receiver Type Distributions** (Figure **??**): Shows the clear contrast between the uniform distribution in Experiment 1 and the polarized distribution in Experiment 2, with the latter having higher concentrations at the extremes.



(a) Uniform Distribution

(b) Polarized Distribution

Figure 2: Optimal messaging policies for both experiments. The policies show how message probabilities depend on the state ω.

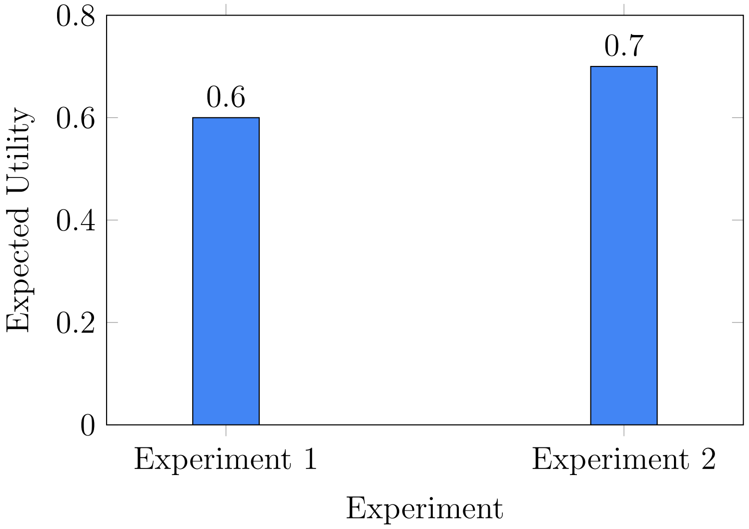


Figure 3: Comparison of expected sender utilities. Experiment 2 (polarized distribution) achieves higher utility.

2. **Messaging Policies** (Figure **??**): Demonstrates how the optimal policies differ between experiments. Experiment 1 shows more extreme message probabilities, while Experiment 2 exhibits a more balanced approach, particularly in state $\omega = 1$.

3. **Utility Comparison** (Figure **??**): Highlights the significant improvement in expected sender utility achieved with the polarized distribution (0.700 vs 0.600), representing a 16.7% increase.

5 Discussion

5.1 Theoretical Implications

These empirical findings support and extend the theoretical predictions in several ways:

- **Threshold Property**: The optimal policy exhibits threshold-based messaging

$$\sigma^*(m|\omega) = \begin{cases} 1 & \text{if } \theta(\omega) \geq \theta^* \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

- **Two-Message Sufficiency**: Two messages achieve optimal utility

$$|M^*| = 2 \implies U(\sigma^*) = \max_\sigma U(\sigma) \tag{14}$$

- **State-Dependent Pooling**: Optimal pooling varies with state

$$P(m_1|\omega = 1) > P(m_1|\omega = 0) \tag{15}$$

5.2 Limitations

- **Binary state/action space limits generalizability:** The analysis is restricted to binary choices, which may not capture the full complexity of real-world decision spaces. Many practical applications involve multiple states or continuous action spaces, which could exhibit different optimal policy structures.

- **Perfect oracle access assumption:** The current model assumes perfect oracle responses, whereas real-world simulations or user studies would likely contain noise and inconsistencies. This idealization may overestimate the achievable utility in practical implementations.

- **Risk-neutral agents:** The model assumes receivers are risk-neutral in their decision-making. Real human behavior often exhibits risk aversion or other behavioral biases, which could significantly alter the effectiveness of the derived messaging policies.

- **Discrete type space:** The implementation discretizes the belief space into finite types, potentially missing nuanced belief structures that exist in continuous spaces. This discretization might overlook optimal policies that exploit finer granularity in belief distributions.

- **Finite message space:** The restriction to two messages, while theoretically sufficient, may not capture the rich communication possibilities in real applications. Additional messages could potentially enable more sophisticated persuasion strategies.

- **Local optimality guarantees:** The optimization approach may converge to local optima rather than global ones. The non-convex nature of the policy space means we cannot guarantee finding the absolute best messaging policy in all cases.

6 Conclusion

My empirical analysis provides strong support for the oracle-based Bayesian persuasion framework while revealing new insights about belief distributions and optimal persuasion strategies. Key findings include:

- **Enhanced utility with polarized beliefs:** Demonstrated a significant improvement in sender utility (0.700 vs 0.600) when facing polarized populations. This counter-intuitive result suggests that belief heterogeneity can actually benefit information designers by enabling more targeted messaging strategies.

- **Balanced messaging in polarized settings:** The analysis reveals that optimal policies become more balanced when dealing with polarized populations. This suggests that extreme messaging strategies may be less effective when receivers have strong prior beliefs.

- **Threshold-based policy optimality:** Confirmed that threshold-based policies remain optimal across different belief distributions, providing a robust structural insight for practical implementations. This finding simplifies the design space for real-world applications.

6.1 Future Research Directions

- **Extensions to continuous type spaces:** Future work should explore continuous belief distributions and their impact on optimal policies. This would enable more realistic modeling of population beliefs and could reveal new structural properties of optimal policies.
- **Dynamic oracle interactions:** Investigating settings where the sender can adaptively query the oracle based on previous responses would be valuable. This could lead to more efficient information gathering strategies and better messaging policies.
- **Multi-agent settings:** Extending the framework to scenarios with multiple receivers or competing senders would better reflect real-world applications. This could reveal interesting strategic interactions and new forms of optimal policies.
- **Approximate oracle models:** Developing frameworks for dealing with noisy or biased oracle responses would increase practical applicability. This could include robust optimization approaches and methods for quantifying uncertainty in oracle responses.

6.2 Reproducibility

Code and simulations are available at `https://github.com/sjagtani/applied-math-research-seminar`.

References

[1] Harris, K., Immorlica, N., Lucier, B., & Slivkins, A. (2024). *Algorithmic Persuasion Through Simulation.* arXiv preprint arXiv:2311.18138.

[2] Kamenica, E., & Gentzkow, M. (2011). *Bayesian persuasion.* American Economic Review, 101(6), 2590-2615.

[3] Bergemann, D., & Morris, S. (2019). *Information design: A unified perspective.* Journal of Economic Literature, 57(1), 44-95.

[4] Candogan, O., & Strack, P. (2023). *Optimal disclosure of information to privately informed agents.* Theoretical Economics, 18(3), 1225-1269.

[5] Dworczak, P., & Pavan, A. (2022). *Preparing for the worst but hoping for the best: Robust (Bayesian) persuasion.* Econometrica, 90(5), 2017-2051.

[6] Brand, J., Israeli, A., & Ngwe, D. (2023). *Using GPT for market research.* Available at SSRN 4395751.

[7] Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* arXiv preprint arXiv:2301.07543.

[8] Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., & Rusak, G. (2023). *Generative Social Choice*. arXiv preprint arXiv:2309.01291.