Consistency and Strategic Disclosure in Model-based Persuasion

Jun Hyun Ji, University of Pittsburgh

I study the problem of persuading a naive Bayesian
by providing a **consistent interpretation** of **selected data**

Motivation

Persuasion is often achieved by providing **models** that

✶ fit the observed data well and provide consistent explanations

e.g., A climate scientist's model matches past weather—persuasive

The model assumes that physical laws vary by season—lose trust

➢ Being consistent is therefore **costly** to persuaders

(1) How costly is Consistency in Model-based Persuasion?

(2) Does it create incentives to frame the data? and How?

✶ Introduce a **"consistency constraint"** in model-based persuasion and study its impact on the scope of persuasion

✶ Endow Sender with **abilities to disclose partial data** and study **optimal data framing** under consistency constraint

Setup

➢ **Unknown state** $\omega \in \Omega = \{G, B\}$ with prior $\mu_0 \in \text{int}(\Delta\Omega)$

➢ **History** $h = (s_1, s_2, \ldots, s_N) \in S^N$ where S is a signal space

➢ **Sub-history** $h_I = (s_i)_{i \in I}$ where $I \subseteq \{1, \ldots, N\}$

➢ **Model** p : distribution $p(\cdot \mid \omega) \in \Delta S^N$ over histories, for each ω

➢ h_I**-fit of model** p measures how likely model p fits sub-history h_I

$$\Pr(h_I \mid p) := \sum_{\omega \in \Omega} \mu_0(\omega) p(h_I \mid \omega)$$

Timeline

Observation → **Persuasion** → **Learning**

Sender privately observes history h

Sender discloses a sub-history h_I & proposes model p

Receiver observes h, updates belief using p, & chooses action

Sender chooses h_I and p to maximize posterior belief about G

$$\mu_h(G \mid p) = \frac{\mu_0(G) p(h \mid G)}{\Pr(h \mid p)}$$

Fit-dominance Constraint (h_I-FD)

$$\Pr(h_I \mid p) \geq \Pr(h_I \mid p_d)$$

where p_d: Receiver's **default model***

Receiver adopts the proposed model if it makes the data he has seen appear *more likely* than his **initial interpretation***

Receiver's Naivete

✶ No prior over models; No ability to generate models

✶ Does not adjust for Sender's incentives

✶ Once persuaded, Receiver forgets his default worldview

Consistency Constraint (IID)

Receiver demands a model such that each signal s_i in history h is IID.

➢ **IID model** π: a distribution $\pi(\cdot \mid \omega) \in \Delta S$ over *signals*, for each ω

Proposed model is adopted if it provides a consistent principle of how each signal arises

(Result 1) Exponential Cost of Consistency: Under full disclosure, Entropy of data diminishes persuasion exponentially

Assume: ❶ History h is fully disclosed ❷ Default model p_d is induced by an IID model π_d

$$\frac{\text{Max. posterior about } G \text{ under (FD)}}{\text{Max. posterior about } G \text{ under (FD) \& (IID)}} = \frac{\mu_{\text{FD}}(G)}{\mu_{\text{FD\&IID}}(G)} = \exp\left(N \cdot \text{entropy}(\hat{\pi}_h)\right) \geq 1$$

where $\hat{\pi}_h$ is the **empirical signal-distribution** of h e.g., $h = (\text{Head, Tail, Tail, Tail}) \Longrightarrow \hat{\pi}_h = (25\%,\ 75\%)$

✶ $1 \leq \exp(N \cdot \text{entropy}(\hat{\pi}_h)) \leq N^N$ *Shannon entropy$(\hat{\pi}_h) = -\sum_{s \in S} \hat{\pi}_h(s) \log \hat{\pi}_h(s)$

✶ $\mu_{\text{FD}}(G) = \dfrac{\mu_0(G)}{\Pr(h \mid p_d)}$ ⋯ (Schwartzstein & Sunderam, 2021, AER) "Level of surprise ↑ $\Longrightarrow$ persuasion ↑"

EXAMPLE

➢ **History** $h = (\text{Head, Tail})$, Prior $\mu_0(G) = 20\%$

➢ **Default**: $p_d(h \mid G) = 25\%$, $p_d(h \mid B) = 25\%$

➢ **Default fit**: $\Pr(h \mid p_d) = 20\% \cdot 25\% + 80\% \cdot 25\% = 25\%$

➢ **Default Posterior**: $\mu_h(G \mid p_d) = 20\%$

PERSUASION WITHOUT IID. (under full disclosure)

➢ Model: $p(h \mid G) = 100\%$, $p(h \mid B) = 6.25\%$

➢ Fit: $\Pr(h \mid p) = 20\% \cdot 100\% + 80\% \cdot 6.25\% = 25\%$ **(FD✓)**

✶ **(FD)** imposes a lower bound on $p(h \mid B)$

➢ Posterior: $\mu_h(G \mid p) = \frac{20\% \cdot 100\%}{25\%} = 80\% = \mu_{\text{FD}}(G)$

PERSUASION WITH IID. (under full disclosure)

➢ iid Model π: $\pi(\text{Head} \mid G) = \pi(\text{Head} \mid B) = 50\%$

✶ **(IID)** imposes an upper bound on $p(h \mid G)$

➢ Posterior: $\mu_{\text{FD\&IID}}(G) = 20\%$

$$\Longrightarrow \frac{\mu_{\text{FD}}(G)}{\mu_{\text{FD\&IID}}(G)} = 4$$

Without (IID) constraint, the posterior belief is 4 times higher.

(Result 2) If history contains two or more mixed signals, then partial data disclosure always grants maximal persuasion

Assume: ❶ Sender can disclose any sub-history
❷ $s_i \neq s_j$ for some i, j in $h = (s_1, .., s_N)$

$\mu_h(G \mid \pi) = 100\%$ is induced by **disclosing any sub-history that excludes at least one signal type**. Sender gains full flexibility in interpreting the undisclosed signals when they are eventually revealed.

e.g., $h = (\text{Head, Tail}) \Longrightarrow$ Disclose $h_I = (\text{Head})$

Propose $\pi(\text{Head} \mid G) = 99\%$ $\pi(\text{Head} \mid B) = 100\%$

$\pi(\text{Tail} \mid G) = 1\%$ $\pi(\text{Tail} \mid B) = 0\%$

$\Longrightarrow$ Tail becomes conclusive evidence for state G when Receiver eventually observes $h = (\text{Head, Tail})$.

✶ Engineering "surprise" to create confirmation bias

Disclosure Problem Given History of N identical signals (Results 3-4)

$$h = \underbrace{\Big(\underbrace{s, s, s, s \ldots, s}_{\text{disclose } K \text{ signals}}, \ldots, s\Big)}_{N \text{ signals}}$$

Sender faces a **TRADE-OFF** between Disclosing **more** signals ($K \uparrow$) *vs.* Disclosing **less** signals ($K \downarrow$) due to interaction betw. **(FD)** and **(IID)** constraints

(Result 3) Discontinuous Disclosure Strategy

Assume: ❶ Sender can disclose $K \in \{1, \ldots, N\}$ signals
❷ Default model p_d is induced by an IID model π_d
❸ $N \geq 2$ is small enough (to avoid triviality)

Optimal disclosure is either $K^* = 1$ or $K^* = N$.

(Result 4) When is $K^* = 1$ optimal?

$K^* = 1 \iff$ Receiver is initially **pessimistic** enough:

➢ Unfavorable prior: $\mu_0(G) < 50\%$

➢ Signal s is initially interpreted as a bad sign:
i.e., **Default model**'s gap $\pi_d(s \mid B) - \pi_d(s \mid G)$ is large enough

If Receiver is initially pessimistic, then disclosing more signals pushes his posterior away from state G, strengthening his initial pessimism and making it harder for Sender to pull it back toward state G.

Reduced Problem Given N Identical Signals

At which K will Receiver accept
the lowest per-signal likelihood in state B?

$$\min_{K \in \{1, \ldots, N\}} \pi(s \mid B)$$

By disclosing K signals,

➢ **(FD)** constrains *the likelihood of K occurrences, in state B*

$$\mu_0(G) \underbrace{p(K \text{ times} \mid G)}_{100\%} + \mu_0(B) p(K \text{ times} \mid B) \equiv \underbrace{\Pr(K \text{ times} \mid \pi_d)}_{\text{Default Fit}}$$

➢ **(IID)** constrains *the per-signal likelihood, in state B*

$$\pi(s \mid B) \equiv p(K \text{ times} \mid B)^{1/K}$$

Trade-off when disclosing more signals

$$\frac{d}{dK} \log \pi(s \mid B) = -\frac{1}{K}\left(\underbrace{\left|\frac{d}{dK} \log p(K \mid B)\right|}_{\text{MB of increasing } K} - \underbrace{\left|\frac{1}{K} \cdot \log p(K \mid B)\right|}_{\text{MC of increasing } K}\right)$$

➢ **MB**: Default fit worsens with more signals $\Longrightarrow$ $p(K \mid B) \downarrow$

➢ **MC**: more signals imply a higher likelihood $\Longrightarrow$ $\pi(s \mid B) \uparrow$

Takeaway

➢ **Demand for consistency** creates strong **incentives to frame data**.

➢ **The dynamics of surprise** influence data framing—not just the ultimate level of surprise—contrast Schwartzstein & Sunderam (2021, AER)

➢ **Heterogeneous** data create extreme incentives to frame data.

➢ **Homogeneous** data create an intense **trade-off** between engineering "surprise" and maintaining consistency.

✶ **Policy implication:** Transparency is not always optimal policy to minimize Sender's influence