INFORMATION AND LEARNING IN ECONOMIC THEORY

(IN PROGRESS AND INCOMPLETE)

ANNIE LIANG
annie.liang@northwestern.edu

Preface

These lecture notes are written for graduate students, although several chapters can also be appropriate for an advanced undergraduate audience (e.g., Chapters 1-4 and Chapter 8). Exercises are labelled U for "undergraduate," G for "graduate," and G^* to suggest that the problem is more involved and/or relies on background knowledge that is not covered in these notes.

Parts of these lecture notes are based on slides by Drew Fudenberg and Xiaosheng Mu. Hershdeep Chopra, Andrei Iakovlev, and Tiago Cardoso Botelho provided very helpful proofreading. I am also grateful to Kyohei Okumura, DJ Thornton, and Jingyi Qiu for comments and corrections. All errors are mine alone. Please feel free to email annie.liang@northwestern.edu with any edits and/or suggestions.

Contents

Part 1

Foundations of Information

Chapter 1

Information Partitions and Knowledge

Section 1.1 introduces the partitional model of information and three definitions of common knowledge. Sections 1.2 presents Aumann (1976)'s result that agents cannot agree to disagree. Section 1.3 presents Rubinstein (1989)'s email game. Section 1.4 defines common p-belief.

We assume a finite state space in Sections 1.1-1.4 to ease exposition, and discuss in Section 1.5 how these results extend more generally.

1.1 Common Knowledge

An unknown state ω takes values in the finite set Ω. Agents $i \in \mathcal{I}$ share a *common prior* that the state ω is distributed according $P \in \Delta(\Omega)$. Each agent i's *information partition* Π_i is a partition of Ω, with the property that for any realization of the state ω, agent i is informed that the state belongs to $\Pi_i(\omega)$.

Assumption 1. *Every partition element has strictly positive probability under the prior; that is, $P(\Pi_i(\omega)) > 0$ for every agent $i \in \mathcal{I}$ and state $\omega \in \Omega$.*

Definition 1.1 (Knowledge). *The set of states at which agent i* knows *the event $A \subseteq \Omega$ to be true is*

$$K_i(A) = \{\omega : \Pi_i(\omega) \subseteq A\}.$$

No agent can think that an event is true if it is not; that is, $K_i(A) \subseteq A$ for every agent i and event A.

Definition 1.2 (Mutual Knowledge). *The set of states at which the event $A \subseteq \Omega$ is* mutual knowledge *is*

$$K(A) = \bigcap_{i \in \mathcal{I}} \{\omega : \Pi_i(\omega) \subseteq A\}$$

i.e., all agents know A to be true.

Example 1.1. Suppose the set of states is $\Omega = \{1,2,3,4,5,6\}$, and there are two agents with information partitions $\Pi_1 = \{\{1,2,3\},\{4,5\},\{6\}\}$ and $\Pi_2 = \{\{1,2\},\{3,4\},\{5\},\{6\}\}$. Let $A = \{3,4,5,6\}$. Then, the set of states at which agent 1 knows A to be true is $K_1(A) = \{4,5,6\}$, the set of states at which agent 2 knows A to be true is $K_2(A) = \{3,4,5,6\}$, and the set of states at which both agents know A to be true is $K(A) = \{4,5,6\}$.

The knowledge operators K_i and K can be applied to events that themselves represent knowledge or mutual knowledge of a state, thus building up higher-order knowledge (agent 1 knows that agent 2 knows that. . .).

Exercise 1.1 (U). *Suppose there are two agents indexed to $i = 1,2$.*

(a) Prove that $K_1(K_2(A)), K_2(K_1(A)) \subseteq K(A)$ for every event $A \subseteq \Omega$.

(b) Provide an example in which $K(A) \not\subseteq K_1(K_2(A))$, demonstrating that even if both players know an event to be true, either can fail to know that the other knows it.

Exercise 1.2 (U). *Prove that $\neg K_i(\neg K_i(A)) = K_i(A)$ for every event $A \subseteq \Omega$ (where $\neg A$ denotes the complement of A.)*

An implicit assumption is made that all agents know the state space Ω and the information partitions $(\Pi_i)_{i\in\mathcal{I}}$. This assumption is less strong than it might initially seem, since we can always redefine states and expand the state space to accommodate uncertainty about other players' partitions, as in the following example.

Example 1.2. Let $\Omega = \{1,2\}$, $\mathcal{I} = \{1,2\}$, and $\Pi_1 = \Pi_2 = \{\{1\},\{2\}\}$. Suppose we want to model the situation where agent 1 has uncertainty over whether agent 2's information is the complete partition Π_2 or the trivial partition $\Pi_2' = \{\{1,2\}\}$. One way to do this is to expand the state space: Define $\widetilde{\Omega} = \Omega \times \{c,t\} = \{\{1,c\},\{1,t\},\{2,c\},\{2,t\}\}$ and revise the agents' information partitions to be

$$\widetilde{\Pi}_1 = \{\{(1,c),(1,t)\},\{(2,c),(2,t)\}\}$$
$$\widetilde{\Pi}_2 = \{\{(1,c)\},\{(2,c)\},\{(1,t),(2,t)\}\}$$

Then, for example, at state $(1,c)$ both agents know $\omega = 1$ to be true, but agent 1 does not know whether agent 2 knows it.

The event A is *common knowledge* at state ω if both agents know it to be true, know the other to know it to be true, ad infinitum. There are at least three equivalent ways to define this.

The First Definition. The most direct approach is to recursively define higher-order levels of knowledge.

Definition 1.3 (Common Knowledge, Definition 1). *For any event $A \subseteq \Omega$, define $\mathscr{A}^1 := \bigcap_{i \in \mathcal{I}} K_i(A)$ to be the set of states at which every agent knows A, and recursively define*

$$\mathscr{A}^k := \bigcap_{i \in \mathcal{I}} K_i(\mathscr{A}^{k-1})$$

for each $k \geq 2$. (For example, $\mathscr{A}^2$ is the set of states at which every agent knows that every agent knows A.) The set of states at which A is common knowledge *is $\mathscr{A}^\infty := \bigcap_{n \geq 1} \mathscr{A}^n$.*

Exercise 1.3 (G). *Consider the informational environment of Example 1.1. Find the smallest value of k with the property that $\mathscr{A}^{k'} = \mathscr{A}^{k'+1}$ for all $k' \geq k$.*

The Second Definition. Alternatively, we can define common knowledge using the meet of the players' information partitions. If two partitions Π and Π' satisfy

$$\Pi'(\omega) \subseteq \Pi(\omega) \quad \forall \omega \in \Omega$$

then we say that Π is a *coarsening* of Π' (corresponding to weakly less information at every state), and Π' is a *refinement* of Π (corresponding to weakly more information at every state). If Π' a coarsening of both partitions Π_1 and Π_2, then it is a *common coarsening* of Π_1, Π_2.

Definition 1.4. *Let $\Pi_1 \wedge \Pi_2$ denote the finest common coarsening of Π_1, Π_2, i.e., the common coarsening of these partitions that is moreover a refinement of every other common coarsening of Π_1 and Π_2.*

Definition 1.5. *For any sequence of information partitions $(\Pi_1, \ldots, \Pi_{|\mathcal{I}|})$, let $\mathscr{P}_2 = \Pi_1 \wedge \Pi_2$, and for each $k > 2$, recursively define $\mathscr{P}_k = \mathscr{P}_{k-1} \wedge \Pi_k$. The* meet *of $(\Pi_1, \ldots, \Pi_{|\mathcal{I}|})$ is $\bigwedge_{i \in \mathcal{I}} \Pi_i \equiv \mathscr{P}_{|\mathcal{I}|}$.*

Exercise 1.4 (G). *Prove that for any sequence of information partitions $(\Pi_1, \ldots, \Pi_{|\mathcal{I}|})$, the meet $\mathscr{P}^n$ is invariant to permutations of players indices.*

Example 1.3. Consider Example 1.1. Stack the two information partitions on top of one another, and suppose an ant is placed on one of the states in an agent's partition (see Figure 1.1).

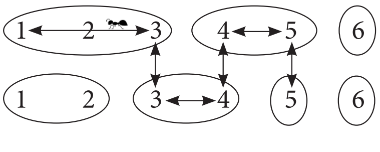

Figure 1.1: $\Pi_1 \wedge \Pi_2(\omega)$ includes all states that an ant seeded at ω can reach.

The ant's movements obey two laws: The ant can move from side to side within an information partition element, and it can jump across the players' information partitions along the same state. The ant's full range of motion when seeded at any state ω then recovers the member of the meet that includes that state. So in this example, we have $\Pi_1 \wedge \Pi_2 = \{\{1, 2, 3, 4, 5\}, \{6\}\}$.

Exercise 1.5 (G). *Formalize the statements in the example above by proving that two points x' and x'' belong to the same element of $\bigwedge_{i\in\mathcal{I}}\Pi_i$ if and only if there is a sequence $(x_0, x_1, x_2, \ldots, x_n, x_{n+1})$, with $x_0 = x'$ and $x_{n+1} = x''$, such that for every $0 \leq m \leq n$, x_m and x_{m+1} belongs to the same element of Π_i for some $i \in \mathcal{I}$.*

Definition 1.6 (Common Knowledge, Definition 2). *An event $A \subseteq \Omega$ is common knowledge at state $\omega \in \Omega$ if $\bigwedge_{i\in\mathcal{I}}\Pi_i(\omega) \subseteq A$.*

Remark 1.1. It is immediate that the set Ω is common knowledge at every $\omega \in \Omega$.

The Third Definition. Our final definition of common knowledge starts from the definition of an *evident* event which, upon its occurrence, is known to all agents.

Definition 1.7 (Evident Events). *The event $A \subseteq \Omega$ is* evident *(or* public*) if $A \subseteq K(A)$.*

Definition 1.8 (Common Knowledge, Definition 3). *The event $A \subseteq \Omega$ is common knowledge at ω if and only if there is an evident event E such that $\omega \in E$ and $E \subseteq K(A)$.*

Exercise 1.6 (G). *Let $\mathcal{I} = \{1, 2\}$. Prove that an event $E \subseteq \Omega$ is evident if and only if it is a union of elements of the meet $\Pi_1 \wedge \Pi_2$.*

These three definitions of common knowledge are equivalent (see for example Monderer and Samet (1989)).

1.2 Agreeing to Disagree

Often we are interested not only in agents' knowledge (which depends only on the agents' information partitions) but also in agents' posterior beliefs (which depend additionally on the prior P). At any state ω and for any event $A \subseteq \Omega$, agent i's posterior probability of event A is pinned down by Bayes' rule (see Section 2.2):

$$P(A \mid \Pi_i(\omega)) = \frac{P(A \cap \Pi_i(\omega))}{P(\Pi_i(\omega))}$$

Our assumption that every partition element has strictly positive prior probability ensures that this expression is well-defined.

One event of interest is the one in which a player's posterior belief takes on a particular value. Fixing an event A and a number $p \in [0, 1]$, define $A_p = \{\omega : P(A \mid \Pi_i(\omega)) = p\}$ to be the set of states at which player i assigns posterior probability p to the event A being true. If player 1 announces that he assigns probability p to A, then all other agents know that the state must belong to A_p.

Example 1.4. Consider the informational environment of Example 1.1 with a uniform prior on Ω, and define the event $A = \{2,3\}$. Agent 2 has four partition elements, $\{1,2\}$, $\{3,4\}$, $\{5\}$, $\{6\}$, and assigns to A a posterior probability of 1/2, 1/2, 0, and 0 (respectively) on these partition elements. So the set of states $A_{1/2}$ at which agent 2 assigns probability 1/2 to event A being true, is $A_{1/2} = \{1,2,3,4\}$.

The following theorem shows that whenever players' posterior beliefs about an event are common knowledge (e.g., because players have publicly announced these beliefs), then these posterior beliefs must be identical. So disagreement cannot be sustained whenever players' beliefs are commonly known.

Theorem 1.1 (Aumann (1976)). *Suppose $\mathcal{I} = \{1,2\}$. Fix any state $\omega \in \Omega$ and event $A \subseteq \Omega$. If it is common knowledge at ω that agent 1 assigns (posterior) probability q_1 to event A, while agent 2 assigns (posterior) probability q_2 to the same event, then $q_1 = q_2$.*

The result is stated for two agents, but the proof below directly extends for an arbitrary finite number of players.

Proof. Let $\mathbf{P}$ be the element of $\Pi_1 \wedge \Pi_2$ that contains ω. Then we can write $\mathbf{P} = \cup_k \mathcal{P}^k$ where $\mathcal{P}^k$ are elements of Π_1. Since the event {agent 1's posterior belief is q_1} is common knowledge at ω, agent 1 must assign probability q_1 to event A at every partition element $\mathcal{P}^k$. So $q_1 = P(A \cap \mathcal{P}^k)/P(\mathcal{P}^k)$ for each k. This implies $q_1 \cdot P(\mathcal{P}^k) = P(A \cap \mathcal{P}^k)$. Summing over each of player 1's partition elements, we have $q_1 \sum_k P(\mathcal{P}^k) = \sum_k P(A \cap \mathcal{P}^k)$. Thus $q_1 \cdot P(\mathbf{P}) = P(A \cap \mathbf{P})$. But repeating the same line of logic for player 2, we obtain $q_2 \cdot P(\mathbf{P}) = P(A \cap \mathbf{P})$. So it must be that $q_1 = q_2$. ■

The following example explains why it is important that players' posterior beliefs are common knowledge and not simply mutual knowledge.

Example 1.5. Let $\Omega = \{1,2,3,4\}$ with a uniform prior, and define $\Pi_1 = \{\{1,2\},\{3,4\}\}$ and $\Pi_2 = \{\{1,2,3\},\{4\}\}$. Choose $A = \{1,4\}$ and $\omega = 2$. Then agent 1 assigns posterior probability 1/2 to A while agent 2 assigns posterior probability 1/3. Moreover, each agent knows one another's posterior probability. But Theorem 1.1 is not violated: agent 2 does not know that agent 1 knows his posterior probability to be $\frac{1}{3}$, so posterior beliefs are mutual knowledge but not common knowledge.

The starting hypothesis of Theorem 1.1—that individuals have common knowledge of one anothers' beliefs—is strong. Geanakoplos and Polemarchakis (1982) show that the same result obtains under a more realistic process: Communication of posterior beliefs converges to common knowledge of identical posterior beliefs, where this convergence occurs in fewer than $n_1 + n_2$ steps with n_i the size of agent i's partition.

Example 1.6. Let $\Omega = \{1,2,3,4,5,6,7,8,9\}$ with all states equally likely. There are two agents, Bob and Carly, with information partitions

$$\Pi_B = \{\{1,2,3\},\{4,5,6\},\{7,8,9\}\}$$

and

$$\Pi_C = \{\{1,2,3,4\},\{5,6,7,8\},\{9\}\}.$$

Suppose the true state is $\omega = 1$, and the agents repeatedly communicate their beliefs about the event $A = \{3,4\}$.

Round 1: Bob's information partition reveals to him that the state belongs to $\{1,2,3\}$, so he assigns posterior probability $1/3$ to the event A. Carly's information partition reveals to her that the state belongs to $\{1,2,3,4\}$, so she assigns posterior probability $1/2$ to the event A. The two agents announce these posterior beliefs.

Round 2: That Bob assigns probability $1/3$ to A reveals to Carly that Bob was either informed that the state belongs to $\{1,2,3\}$ or informed that the state belongs to $\{4,5,6\}$. But Carly already knew in round 1 that these were the two partition elements that Bob might have been informed of (since she knew the state to be either 1, 2, 3, or 4), and so there is no information for her in this announcement. That Carly assigns probability $1/2$ to A reveals to Bob that agent 2 knows $\{1,2,3,4\}$, but again Bob knew this in round 1. So both agents' posterior beliefs are unchanged. They again announce $1/3$ and $1/2$.

And now something interesting happens. That Bob sticks to his original belief of $1/3$ tells Carly that Bob must have observed $\{1,2,3\}$. If instead Bob observed $\{4,5,6\}$, then upon hearing that Carly's belief was $1/2$ (and thus learning that Carly observed $\{1,2,3,4\}$), Bob would have deduced that the state was 4 with certainty, and hence revised his posterior belief of A to 1. So Carly now knows that the state is in $\{1,2,3\}$ and shares Bob's posterior belief, $\frac{1}{3}$. The two agents' beliefs have converged, and it is straightforward to show that these beliefs will not move after subsequent communication.

Although agents' beliefs must converge, the belief that they converge to need not be the belief that agents would have held had they pooled their information:

Example 1.7. Let $\Omega = \{1,2,3,4\}$ with each state equally likely. Agents' partitions are given by $\Pi_1 = \{\{1,2\},\{3,4\}\}$ and $\Pi_2 = \{\{1,3\},\{2,4\}\}$. Let $\omega = 1$ and $A = \{1,4\}$. Both posteriors are $1/2$ and the process of belief revision converges in one step. But had agents shared their information, they would have learned that $\omega \in \{1,2\}$ and also $\omega \in \{1,3\}$, leading to a (common) posterior belief that A is true with probability 1.

1.3 The Email Game

Common knowledge assumptions appear frequently in analyses of strategic environments; for example, payoffs are assumed to be common knowledge in any complete-information game. Do strategic predictions made under an assumption of common knowledge approximately hold when we relax the assumption of common knowledge? Rubinstein (1989)'s email game shows that for one formalism of what "almost common knowledge" means, the answer is no: Strategic predictions can change discontinuously when we move from common knowledge to almost common knowledge.

In this game, two agents each choose an action from $\{A, B\}$. There are two possible payoff matrices indexed to $\{a, b\}$ (depicted below with a on the left and b on the right). The agents share a common prior that assigns probability $1 - p > \frac{1}{2}$ to the matrix indexed to a.

$$\begin{array}{ccc} & A & B \\ A & M, M & 0, -L \\ B & -L, 0 & 0, 0 \end{array} \qquad\qquad \begin{array}{ccc} & A & B \\ A & 0, 0 & 0, -L \\ B & -L, 0 & M, M \end{array}$$

We assume throughout that $L > M > 0$. Thus (A, A) yields higher payoffs for both agents when the payoff parameter is a while (B, B) yields higher payoffs when the payoff parameter is b. The action A is "safe," in that it never yields a negative payoff.

Communication Protocol. Both players have an automated email server, which is the only means by which the players can communicate. Agent 1 is informed of the payoff parameter. If (and only if) the parameter is b, agent 1's email server automatically sends an email to agent 2 announcing that the parameter is b. All emails are independently lost with probability $\varepsilon > 0$, so the agents' email servers are set up to automatically send back confirmations that emails have been received, and confirmations of confirmations, etc. Each agent i's type is the number of emails that agent i's computer sends, which is privately known to agent i.

In the special case $T_1 = T_2 = \infty$, there is common knowledge that the parameter is b. But if for example $T_1 = 2$, then agent 1 knows the parameter is b, and agent 1 knows that agent 2 knows that the parameter is b, but agent 1 does not know that agent 2 knows that agent 1 knows that agent 2 knows that the parameter is b. In general, so long as T_1 and T_2 are finite, then higher-order knowledge of parameter b must break down at some stage.

Remark 1.2. In the partitional framework of Section 1.1, we would model this information environment as follows: The state space is

$$\Omega = \{(a,0,0), (b,1,0), (b,1,1), (b,2,1), (b,2,2), \dots\}$$

and agents' information partitions are given by

$$\begin{aligned} \Pi_1 &= \{\{(a,0,0)\}, \{(b,1,0), (b,1,1)\}, \{(b,2,1), (b,2,2)\}, \dots\} \\ \Pi_2 &= \{\{(a,0,0), (b,1,0)\}, \{(b,1,1), (b,2,1)\}, \dots\} \end{aligned}$$

where, for example, $T_1 = 0$ reveals to player 1 the partition element $\{(a,0,0)\}$, while $T_2 = 0$ reveals to player 2 the partition element $\{(a,0,0), (b,1,0)\}$.

Proposition 1. *There is a unique Bayesian Nash equilibrium in which agent* 1 *plays A when the payoff parameter is a. In this equilibrium, both agents play A independently of the number of messages sent.*

Proof. Let $s_i : T_i \to \Delta(\{A, B\})$ denote player i's equilibrium strategy. By assumption, $s_1(0) = A$. We will show that also $s_2(0) = A$. Agent 2 of type $T_2 = 0$ knows that either agent 1's first message was never sent (the state is $(a, 0, 0)$), or agent 1's first message was sent but lost (the state is $(b, 1, 0)$). Unconditionally, the probabilities of these states are $(1 - p)$ and $p\varepsilon$. Conditional on $T_2 = 0$, agent 2 assigns a posterior probability of $\frac{1-p}{1-p+p\varepsilon}$ to $(a, 0, 0)$, a posterior probability of $\frac{p\varepsilon}{1-p+p\varepsilon}$ to $(b, 1, 0)$ and zero probability to all other states.

So agent 2's expected payoff from playing A is at least

$$M \cdot \left(\frac{1-p}{1-p+p\varepsilon}\right) + 0 \cdot \left(\frac{p\varepsilon}{1-p+p\varepsilon}\right) \tag{1.1}$$

while agent 2's expected payoff from playing B is no more than

$$(-L) \cdot \left(\frac{1-p}{1-p+p\varepsilon}\right) + M \cdot \left(\frac{p\varepsilon}{1-p+p\varepsilon}\right). \tag{1.2}$$

Since $1 - p > \frac{1}{2}$ and $L > M$ by assumption, (1.1) strictly exceeds (1.2), and so agent 2's strategy must satisfy $s_2(0) = A$.

Now suppose $s_i(T_i) = A$ for $i = 1, 2$ and all $T_i < t$. We'll argue that $s_1(t) = s_2(t) = A$. Suppose first that agent 1's computer sends t emails exactly, i.e., $T_1 = t$. Since agent 1's computer did not send a $(t + 1)$-th email, it must either be that agent 1's t-th message was lost (the state is $(b, t, t - 1)$), or that agent 1's t-th message was received, but its confirmation was lost (the state is (b, t, t)). Agent 2's posterior belief conditional on $T_1 = t$ then assigns probability $z := \frac{\varepsilon}{\varepsilon+(1-\varepsilon)\varepsilon} > \frac{1}{2}$ to $(b, t, t - 1)$ and probability $1 - z$ to (b, t, t). So the expected payoff to playing B is $z(-L) + (1 - z)(M) < 0$, while the payoff to playing A is zero. We conclude that agent 1's strategy must satisfy $s_1(t) = A$, with nearly identical reasoning yielding $s_2(t) = A$. ∎

This result shows a sharp discontinuity in strategic predictions at common knowledge. That is, (B, B) is an equilibrium when agents have common knowledge of the payoff parameter b, but fails to be an equilibrium when players have knowledge of b to arbitrarily high (finite) orders.

Whether this result is surprising depends on how natural we consider the relaxation of common knowledge to be. Rubinstein (1989) argues that "high T_i" is intuitively like common knowledge. Another view is that these are substantially different, since for arbitrarily small but strictly positive ε the informational model is the one described in Remark 1.2, but for $\varepsilon = 0$ (corresponding to common knowledge of the state) the set of states with positive ex-ante probability is $\Omega = \{(a, 0, 0), (b, \infty, \infty)\}$ and the agents' information partitions are complete. So there is a discontinuity in the informational environments as $\varepsilon \to 0$, and in this sense small ε may be quite unlike $\varepsilon = 0$.

1.4 (Common) p-Belief

We now consider an alternative approach to formalizing almost common knowledge, which defines common "almost-knowledge" in contrast to the above "almost-common" knowledge.

Definition 1.9. *For any $p \in [0,1]$, say that agent i p*-believes *A at ω if $P(A \mid \Pi_i(\omega)) \geq p$. The set of states at which agent i p-believes A is*

$$\mathcal{B}_i^p(A) = \{\omega : P(A \mid \Pi_i(\omega)) \geq p\}.$$

Remark 1.3. Is the case $p = 1$ equivalent to knowledge? Suppose $\Omega = \{1,2,3\}$ and the prior is $P = (0, 1/2, 1/2)$. Agent 1's partition is $\{\{1,2\},\{3\}\}$ while agent 2's partition is $\{\{1\},\{2\},\{3\}\}$. The state is $\omega = 2$. Then according to Definition 1.1, agent 2 knows $\{2\}$ but agent 1 does not, while according to Definition 1.9, both agents have 1-belief of $\{2\}$. Whether knowledge and 1-belief represent distinct modes of understanding is an interesting philosophical question, but we will not have more to say on it here.

The following construction of *common p-belief*, due to Monderer and Samet (1989), is parallel to Definition 1.3 for common knowledge.

Definition 1.10 (Common p-Belief). *For any $p \in [0,1]$ and event $A \subseteq \Omega$, define $\mathscr{A}^1 = \bigcap_{i \in \mathcal{I}} \mathcal{B}_i^p(A)$ to be the set of states at which every agent p-believes A to be true, and recursively define $\mathscr{A}^k = \bigcap_{i \in \mathcal{I}} \mathcal{B}_i^p(\mathscr{A}^{k-1})$ for every $k \geq 2$. Then A is common p-belief at the set of states $\mathscr{A}^\infty = \cap_{n \geq 1} \mathscr{A}^n$.*

We can also define common p-belief by generalizing the definition of an evident event (Definition 1.7) to events that are evident p-belief.

Definition 1.11. *For any $p \in [0,1]$, the event $A \subseteq \Omega$ is* evident p-belief *if $A \subseteq \bigcap_{i \in \mathcal{I}} \mathcal{B}_i^p(A)$.*

Definition 1.12. *For any $p \in [0,1]$, the event $A \subseteq \Omega$ is common p-belief at ω if there exists an evident p-belief event E such that*

$$\omega \in E \subseteq \bigcap_{i \in \mathcal{I}} \mathcal{B}_i^p(A).$$

Definitions 1.10 and 1.12 are introduced in Monderer and Samet (1989) and shown to be equivalent.

Exercise 1.7 (G*). *Consider the email game of Rubinstein (1989). Let P denote the common prior on Ω (as defined in Remark 1.2), and define $\mathscr{C}^p$ to be the event that agents have common p-belief in parameter b. For each $\varepsilon \geq 0$, let*

$$\overline{p}(\varepsilon) = \sup_{p \in [0,1]} \{p : P(\mathscr{C}^p) > 0\}$$

be the supremum of the set of values of p such that $\mathscr{C}^p$ has positive ex-ante probability. Is $\overline{p}(0)$ equal to the limit of $\overline{p}(\varepsilon)$ as $\varepsilon \to 0$? Discuss your answer.

1.5 General State Spaces

To show that the preceding insights do not require assumption of a finite state space, we now briefly discuss two generalizations of these ideas. In each case, we begin with a probability space (Ω, Σ, P) where Ω is a set of states endowed with σ-algebra Σ, and $P : \Sigma \to [0,1]$ is a probability measure.

The first generalization. Let each information partition Π_i be a partition of Ω, where we require that each partition element is Σ-measurable and has strictly positive measure under P (see e.g., Monderer and Samet (1989)). All of the above definitions and proofs generalize as stated.

The second generalization. Alternatively, we might model each agent i's information as a sub σ-algebra of Σ, denoted by Π_i.[1] One foundation for this approach (which we will examine in detail in subsequent chapters) is that each agent i privately observes a random variable $X_i : \Omega \to \mathbb{R}$ that is measurable with respect to Σ. In this case, each agent i's σ-algebra is $\sigma(X_i)$, the σ-algebra generated by X_i, which is indeed coarser than Σ.

The definition of knowledge can be extended as follows.

Definition 1.13. *Agent i* knows *the event $A \in \Sigma$ to be true at ω if there exists some $B \in \Pi_i$ such that $\omega \in B \subseteq A$.*

Common knowledge cannot in general be iteratively constructed (à la Definition 1.3) using this definition of K_i, since the set of states at which agent i knows A to be true may not be Σ-measurable. Nevertheless, similar to Definition 1.6, we can define $\bigwedge_{i \in \mathcal{I}} \Pi_i$ to be the finest common coarsening of the σ-algebras $\Pi_1, \ldots, \Pi_n$, and say that an event A is common knowledge at ω if there is an element A of $\bigwedge_{i \in \mathcal{I}} \Pi_i$ such that $\omega \in A$. We can also generalize Definition 1.8 as follows:

Definition 1.14. *The event $A \in \Sigma$ is* evident *if $A \in \Pi_i$ for every $i \in \mathcal{I}$, i.e., A belongs to every agent's σ-algebra.*

Definition 1.15. *The event $A \in \Sigma$ is common knowledge at state ω if there is an evident event E such that $\omega \in E$ and $E \subseteq A$.*

Theorem 1.1 can also be generalized, although the previous proof does not extend (for example, there is no longer guaranteed to be a unique element of $\Pi_1 \wedge \Pi_2$ that contains ω).

Proposition 2. *Let $X \in \mathcal{L}^1(\Omega, \Sigma, P)$, and define $Y = \mathbb{E}(X \mid \Pi_1)$, $Z = \mathbb{E}(X \mid \Pi_2)$. If it is common knowledge that $Y = y$ and $Z = z$ at a state ω with strictly positive probability, then it must be that $y = z$.*

Proof. If it is common knowledge that $Y = y$ and $Z = z$, there must exist an event $E \in \Pi_1 \cap \Pi_2$, where Y takes the constant value y on E, and Z takes the constant value z on E. Let $\mathbb{1}_E$ denote the indicator variable that takes value 1 on E. Then

$$\begin{aligned} y \cdot P(E) &= \mathbb{E}(Y \mathbb{1}_E) \\ &= \mathbb{E}(X \mathbb{1}_E) \\ &= \mathbb{E}(Z \mathbb{1}_E) = z \cdot P(E) \end{aligned}$$

[1] That is, Π_i is a σ-algebra and $\Pi_i \subseteq \Sigma$.

using in the second and third equalities that Y and Z are conditional expectations of X. Since $P(E) > 0$ (by assumption that $\omega \in E$ has strictly positive probability), it follows that $y = z$ as desired. ■

This result is in fact more general than Theorem 1.1, nesting the previous result as a special case when we choose X to be an indicator function on some set.

Exercise 1.8 (G*). *Generalize Proposition 2 by demonstrating that the conclusion still holds if we assume that there is a measurable set of states $B \subseteq \Omega$ with strictly positive probability, where at every $\omega \in B$ it is common knowledge that $Y = y$ and $Z = z$.*

1.6 Additional Exercises

Exercise 1.9 (G). *Two spies in an underground organization are stationed at remote locations. Each spy privately observes whether the coast is clear at their location. The spies share a common prior that the coast is clear at each location independently with probability $1/2$.*

Communication protocol. *The spies communicate by email with a third party electronic server at their home base. If and only if the coast is clear at a spy's location, that spy's computer will automatically send a message to the home base with the information that the coast is clear.*

If the home base electronic server receives information from both spies indicating that the coast is clear, then it will automatically send a message to both spies indicating that it has received both messages. (Otherwise, it will send no messages.) As these are dangerous times, each message has only a $1 - \varepsilon$ chance of being received (again independent). If either spy receives a message from the home base, that spy will send a reply to the home base confirming receipt. The reply is lost with probability ε, independently of everything that's happened before. So on and so forth. Everything stated above is common knowledge.

Each spy observes the number of messages he has sent, and chooses an action in $\{A, B\}$. If the coast is clear at both locations, then payoffs are given by the ***right*** *matrix below, and otherwise payoffs are given by the* ***left*** *matrix below.*

$$\begin{array}{c|cc} & A & B \\ \hline A & M,M & 0,-L \\ B & -L,0 & 0,0 \end{array} \qquad\qquad \begin{array}{c|cc} & A & B \\ \hline A & 0,0 & 0,-L \\ B & -L,0 & M,M \end{array}$$

The payoff parameters satisfy $L > 3M > 0$.

(a) *Prove the following analogue of Rubinstein (1989)'s result: Let $T_1 = \mathbb{Z}_+$ and $T_2 = \mathbb{Z}_+$ denote the two players' type spaces. There is a unique pure-strategy equilibrium in which both players choose A when the coast is* ***not*** *clear at their location, i.e. $s_1(0) = s_2(0) = A$. In this equilibrium, players choose A for any number of messages sent, i.e. $s_i(t) = A$ for both players i and all $t \in T_i$.*

(b) Suppose instead that $L = 2$ while $M = 1$, and demonstrate that the result in Part (a) no longer holds by finding some $\varepsilon > 0$ and a pair of strategies (s_1, s_2) that constitute a pure-strategy Bayesian Nash equilibrium, where $s_1(0) = s_2(0) = A$ and $s_i(t) = B$ for some player i and type $t \in T_i$.

Exercise 1.10 (G*). *Let $X \in \mathcal{L}^1(\Omega, \Sigma, P)$, and define $Y = \mathbb{E}(X \mid \Pi_1)$, $Z = \mathbb{E}(X \mid \Pi_2)$. Prove that if it is common knowledge that $Y \in A$ and $Z \in B$ at a state ω with strictly positive probability, then it must be that $A \cap B \neq \emptyset$.*

Chapter 2

Bayesian Updating and Beliefs

Section 2.1 introduces the canonical Bayesian framework and the definition of a signal. Section 2.2 reviews Bayes' rule and key properties of Bayesian posteriors. Section 2.3 provides closed-form expressions for posterior beliefs in the special case of Bayesian updating to normal signals, with applications.

2.1 Preliminaries

There is a set of *parameters* Θ endowed with a σ-algebra Σ. An agent has a *prior* $p \in \Delta(\Theta)$, where $\Delta(\Theta)$ denotes the set of Σ-measurable probability measures on Θ. The prior describes the agent's belief at an "ex-ante" stage in the absence of any information, where what is ex-ante is understood in the context of a specific model.

The focus of this chapter is the object that we will call an *information structure*, *experiment*, or a *signal*, which can be formalized in either of several ways:

(a) We can define the signal to be a mapping $\sigma : \Theta \to \Delta(S)$ from the set of parameters to distributions over a set of signal realizations S. See for example de Oliveira (2019).

(b) We can define a signal to be an $(S, \mathcal{S})$-valued random variable X on an underlying probability space (Ω, Σ, P), where $\Omega = \Theta \times E$ for some set E. For example, we might define the signal to be $X = \theta + \varepsilon$ for an E-valued noise term ε that is independent of θ, as we do in Section 2.3.

(c) We can define a signal S to be a finite partition of $\Omega = \Theta \times [0, 1]$, whose elements are non-empty and measurable with respect to the Lebesgue σ-algebra on Ω. Conditional on parameter θ, the probability of observing $s \in S$ is the Lebesgue measure of $\{x \in [0, 1] \mid (\theta, x) \in s\}$. See for example Frankel and Kamenica (2019).

Remark 2.1. It is straightforward to see that the first two formalisms nest one another when all the relevant sets are finite. Suppose we are given a prior $p \in \Delta(\Theta)$ and a signal $\sigma : \Theta \to \Delta(S)$. Define the expanded state space to be $\Omega = \Theta \times S$ and let $P(\theta, s) = p(\theta)\sigma(s \mid \theta)$. Then the random variable

$X : \Omega \to S$ satisfying $X(\theta, s) = s$ is equivalent to σ in the sense that posterior beliefs about θ are the same whether we condition on the realization of X or the realization of $\sigma(\theta)$. In the other direction, if we start with a random variable $X : \Theta \times E \to S$ and a distribution $P \in \Delta(\Theta \times E)$, then we can define $\sigma : \Theta \to \Delta(S)$ to satisfy $\sigma(s \mid \theta) = P(X^{-1}(s) \mid \theta)$. The formalism in (c) is a special case of (b), where $E = [0,1]$, the random variable $X : \Omega \to S$ maps each ω into the partition element of S to which it belongs, and the probability distribution P is the Lebesgue measure.

Example families of signals include:

Example 2.1 (Aumann (1976)'s Partitional Information Structures). For each agent i, let Π_i be a finite partition of Θ into measurable elements of strictly positive measure. Index these partition elements to $S = \{1, \ldots, n\}$ where n is the size of Π_i. Then let σ map each θ with probability 1 to the index of the partition element to which θ belongs.

Example 2.2 (Finite Information Structures). Suppose $|\Theta|, |S| < \infty$. Then we can express σ as a $|\Theta| \times |S|$ matrix where (1) all entries are nonnegative, and (2) all rows sum to 1. For example, suppose a drug is either good (g) or bad (b). The drug is administered to a patient who is either cured (C) or not (N). The patient is cured with probability 3/4 if the drug is good and with probability 1/4 if the drug is bad. Then $\Theta = \{g, b\}$ and $S = \{C, N\}$ and the information structure is

$$\begin{array}{ccc} & C & N \\ g & 3/4 & 1/4 \\ b & 1/4 & 3/4 \end{array}$$

with each row depicting the probability over the signal realizations in the associated state.

Example 2.3 (Gaussian Information). The signal is $X = \theta + \varepsilon$, where $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$, and $\theta \perp\!\!\!\perp \varepsilon$.

2.2 Posterior Beliefs

2.2.1 Bayes' Rule

The agent updates his prior to the realization of the signal using Bayes' rule.

Definition 2.1 (Bayes' Rule, Finite Case). *Suppose $|\Theta| < \infty$. Fix any distribution $P \in \Delta(\Theta)$ and any events $A, B \subseteq \Theta$ where $P(A), P(B) > 0$. Then*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.1}$$

Remark 2.2. Rather than memorizing this formula, it is easier to remember that by the law of total probability, we can rewrite $P(A \cap B)$ as $P(A \mid B)P(B)$ or as $P(B \mid A)P(A)$, so

$$P(A \mid B)P(B) = P(B \mid A)P(A).$$

Dividing through by $P(B)$ yields (2.1).

Remark 2.3. Applying (2.1) twice for the pairs of events (A, E) and (B, E), we have

$$\frac{P(A \mid E)}{P(B \mid E)} = \frac{P(E \mid A)}{P(E \mid B)} \cdot \frac{P(A)}{P(B)}$$

so the relative conditional probabilities of events A and B is determined by their relative probabilities under the prior, $\frac{P(A)}{P(B)}$, and the *likelihood ratio* of E under A and B, $\frac{P(E|A)}{P(E|B)}$. *Base-rate neglect* is the tendency to falsely equate $\frac{P(A|E)}{P(B|E)}$ with $\frac{P(E|A)}{P(E|B)}$, neglecting the prior distribution. This can lead to compelling but inaccurate statistical conclusions.

For example, suppose $\Omega = \{p, n\}$, where $\omega = p$ indicates that an individual is positive for a medical condition while $\omega = n$ indicates that the individual is negative, with $P(\omega = p) = 0.01$. Let $X \in \{+, -\}$ be the outcome of a test where $P(X = + \mid \omega = p) = 0.95$ and $P(X = + \mid \omega = n) = 0.05$. Since the likelihood of observing $X = +$ is much higher when the individual has the condition than when he does not—indeed, the likelihood ratio is $\frac{P(X=+|\omega=p)}{P(X=+|\omega=n)} = 19$—it is tempting to conclude from a positive test result that the individual has the condition. But correctly applying Bayes' rule yields that $\frac{P(\omega=p|X=+)}{P(\omega=n|X=+)} < 1$; that is, even with a positive test it is more likely that the individual is negative for the condition.

A useful rewriting of Bayes' rule is

$$P(\theta \mid X = x) = \frac{P(X = x \mid \theta)P(\theta)}{\sum_{\theta' \in \Theta} P(X = x \mid \theta')P(\theta')} \quad \forall \theta \in \Theta \tag{2.2}$$

where the conditional distribution $P(\cdot \mid X = x)$ is precisely the agent's posterior belief upon observing $X = x$.

Example 2.4. A drug is either effective ($\theta = A$) or not ($\theta = B$), where the prior probability that the drug is effective is $p \in (0, 1)$. The signal is

$$\begin{array}{ccc} & a & b \\ A & q & 1-q \\ B & 1-q & q \end{array}$$

for some $q \in (0, 1)$. Then upon observing a, the agent assigns to $\theta = A$ a posterior probability of

$$\frac{pq}{pq + (1-p)(1-q)} = \frac{1}{1 + \frac{1-p}{p}\left(\frac{1-q}{q}\right)}$$

which exceeds the prior belief of p if and only if $q \geq \frac{1}{2}$.

More generally, when θ and X are (not necessarily finite-valued) random variables with densities f_θ and f_X and conditional densities $f_{\theta|X=x}$ and $f_{X|\theta=t}$, then the posterior belief given $X = x$ is

$$f_{\theta|X=x}(t) = \frac{f_{X|\theta=t}(x) f_\theta(t)}{\int_{\theta' \in \Theta} f_{X|\theta=t'}(x) f_\theta(t') dt'} \quad \forall t \in \Theta. \tag{2.3}$$

Somewhat more generally, we may suppose that the joint distribution of (θ, X) is such that for every realization x of X, there is a (measurable) function q_x satisfying

$$q_x(A) = \mathbb{E}(\mathbb{1}_A \mid X = x) \quad \text{for all events } A \subseteq \Theta$$

Then this q_x is the posterior belief.

2.2.2 Bayes' Plausibility

Outside of special cases (such as the one we will cover in Section 2.3), posterior beliefs often cannot be expressed in closed-form. Nevertheless, there are certain properties they must satisfy. One important property is that beliefs are a martingale, i.e., the expected posterior is equal to the prior. Intuitively, if you expect to change your mind given more information, then why haven't you done so already?

Fact 2.1 (Beliefs are a martingale.). *Let $p \in \Delta(\Theta)$ denote the agent's prior belief, and choose any event A. Then the posterior probability assigned to this event conditional on the realization of random variable X is $\mathbb{E}(\mathbb{1}_A \mid X)$. By the law of iterated expectations,*

$$\mathbb{E}(\mathbb{E}(\mathbb{1}_A \mid X)) = \mathbb{E}(\mathbb{1}_A)$$

so the expected posterior probability of A is equal to the prior probability of A. Since the event A was arbitrarily chosen, we can conclude that the expected posterior belief is equal to the prior belief. (In the case of a finite state space Θ, choosing $A = \{\theta\}$ yields $\mathbb{E}(p(\theta \mid X)) = p(\theta)$ for every θ.)

Since any signal X induces a distribution $\tau \in \Delta(\Delta(\Theta))$ over posterior beliefs, Fact 2.1 implies that this distribution must average to the prior.

Definition 2.2. *Fixing a prior $p \in \Delta(\Theta)$, say that a distribution of posteriors τ is* Bayes plausible *if*

$$\int_{\Delta(\Theta)} q d\tau(q) = p$$

i.e. the expected posterior is equal to the prior. We'll use

$$\mathcal{T}(p) \equiv \left\{ \tau \in \Delta(\Delta(\Theta)) \mid \int q d\tau(q) = p \right\}$$

to denote the set of Bayes plausible posterior distributions given prior p.

Exercise 2.1 (U). *The state space is $\Theta = \{\theta_1, \theta_2\}$ and the prior is $(\mu, 1-\mu)$ for some $\mu \in [0,1]$. The signal structure is*

	s_1	s_2
θ_1	p	$1-p$
θ_2	q	$1-q$

where $p, q \in [0,1]$. What is the distribution over posterior beliefs induced by this signal structure? Verify that the expected posterior belief is equal to the prior belief.

Not only are we guaranteed that any signal induces a Bayes-plausible distribution over posterior beliefs, but also any Bayes-plausible distribution over posterior beliefs can be induced by some signal.

Definition 2.3. *For any signal $X \sim P_X$, let $\tau_X \in \Delta(\Delta(\Theta))$ satisfy $\tau_X(q) = P_X(\{x : q_x = q\})$. Say that $\tau \in \Delta(\Delta(\Theta))$ is* induced by X *if $\tau = \tau_X$.*

Proposition 3. *Suppose the prior p belongs to the interior of the set $\Delta(\Theta)$. Then every Bayes-plausible distribution $\tau \in \mathcal{T}(p)$ is induced by some signal X.*

The proof (demonstrated in Kamenica and Gentzkow (2011) and Shmaya and Yariv (2016) among others) proceeds by construction. For any distribution τ, index the distinct posterior beliefs in the support of τ to be $\{q_x\}_{x \in \mathcal{X}}$, where $\mathcal{X}$ may not be finite. Then define $\sigma : \Theta \to \Delta(\mathcal{X})$ to satisfy

$$\sigma(x \mid \theta) = \frac{q_x(\theta)\tau(q_x)}{p(\theta)} \tag{2.4}$$

We have constructed a signal σ whose realizations x are identified with posterior beliefs q_x, where the conditional distribution over signal realizations mimics Bayes' rule $p(x \mid \theta) = \frac{p(\theta|x)p(x)}{p(\theta)}$, setting $q_x(\theta) = p(\theta \mid x)$ and $\tau(q_x) = p(x)$. This is a valid signal structure since

$$\int_{\mathcal{X}} \sigma(x \mid \theta) dx = \int_{\mathcal{X}} \frac{q_x(\theta)\tau(q_x)}{p(\theta)} dx = 1$$

by (2.4) and the definition of Bayes-plausibility. Moreover,

$$\frac{\sigma(x \mid \theta)p(\theta)}{\int_\Theta \sigma(x \mid \theta)p(\theta)d\theta} = \frac{\sigma(x \mid \theta)p(\theta)}{\tau(q_x)\int_\Theta q_x(\theta)d\theta} = \frac{\sigma(x \mid \theta)p(\theta)}{\tau(q_x)} = q_x(\theta)$$

so $q_x(\cdot)$ is precisely the posterior belief when updating to the signal σ.

Thus the probability that the posterior belief is q_x is exactly the probability that the realization of the constructed signal σ is x, so τ is induced by σ as desired.

Exercise 2.2 (U). *Suppose the prior is over $\Theta = \{\theta_1, \theta_2\}$ is $(1/3, 2/3)$. Provide a set S and a signal structure $\sigma : \Theta \to \Delta(S)$ that induces the belief (0,1) with probability 1/3, and the belief (1/2,1/2) with probability 2/3.*

Together, Fact 2.1 and Proposition 3 imply:

Corollary 2.1. *Fix any prior belief $p \in Int(\Delta(\Theta))$. Then a distribution over posteriors $\tau \in \Delta(\Delta(\Theta))$ is induced by some signal if and only if it is Bayes-plausible, i.e., $\tau \in \mathcal{T}(p)$.*

2.2.3 Application of Bayes' Rule: Incompatibility of Fairness Definitions

Here we take a detour to demonstrate the power of Bayes' rule. Individuals in a population are each described by a covariate vector $C \in \mathcal{C}$, a group membership $G \in \{g_1, g_2\}$, and a type $\theta \in \{0,1\}$. For example, we might interpret θ as the individual's creditworthiness (whether the individual would pay back a loan if approved), G as a demographic group, and C as the individual's credit history. Across individuals, the random vector (C, G, θ) is distributed according to P, and we use $p_g = P(\theta = 1 \mid G = g)$ for the base rate of $\theta = 1$ in each group g. A *scoring rule* is any mapping $S : \mathcal{C} \to \{0,1\}$ that predicts the type given the covariate vector.

Definition 2.4 (Equality of False Positives). *A scoring rule S has equal false positive rates if*

$$P(S = 1 \mid \theta = 0, G = g_1) = P(S = 1 \mid \theta = 0, G = g_2)$$

In words, the probability of being incorrectly assessed to pay back the loan is independent of group membership. Equivalently: $S \perp\!\!\!\perp G \mid \theta = 0$, i.e., the score is conditionally independent of group membership given type $\theta = 0$.

Definition 2.5 (Equality of False Negatives). *A scoring rule S has equal false negative rates if*

$$P(S = 0 \mid \theta = 1, G = g_1) = P(S = 0 \mid \theta = 1, G = g_2)$$

In words, the probability of being incorrectly assessed to not pay back the loan is independent of group membership. Equivalently: $S \perp\!\!\!\perp G \mid \theta = 1$, i.e., the score is conditionally independent of group membership given type $\theta = 1$.

Definition 2.6 (Calibrated). *A score S is calibrated if for each $s \in \{0,1\}$,*

$$P(\theta = 1 \mid S = s, G = g_1) = P(\theta = 1 \mid S = s, G = g_2)$$

In words, among those assessed to pay back the loan (or, to not pay back the loan), the probability of paying back the loan is independent of group membership. Equivalently: $\theta \perp\!\!\!\perp G \mid S$, i.e., type is independent of group membership conditional on the score.

The following impossibility result demonstrates that (outside of edge cases) these fairness criteria cannot be simultaneously satisfied.

Proposition 4 (Kleinberg, Mullainathan and Raghavan (2017),Chouldechova (2017)). *Suppose $p_{g_1} \neq p_{g_2}$. Then no scoring rule S can simultaneously satisfy calibration, equal false positive rates, and equal false negative rates.*

Proof. Choose either group g and define $FP_g = P(S = 1 \mid \theta = 0, G = g)$, $FN_g = P(S = 0 \mid \theta = 1, G = g)$, and $PPV_g = P(\theta = 1 \mid S = 1, G = g)$. We'll show that these quantities are related by the following identity:

$$FP_g = \frac{p_g}{1 - p_g} \times \frac{1 - PPV_g}{PPV_g} \times (1 - FN_g). \tag{2.5}$$

To simplify notation, let Q denote the joint distribution over (C, G, S) after conditioning on $G = g$. Then, expanding (2.5), we have

$$Q(S = 1 \mid \theta = 0) = \frac{Q(\theta = 1)}{Q(\theta = 0)} \times \frac{Q(\theta = 0 \mid S = 1)}{Q(\theta = 1 \mid S = 1)} \times Q(S = 1 \mid \theta = 1)$$

Multiplying both sides by $Q(\theta = 0)$ and applying Bayes' rule,

$$Q(S = 1, \theta = 0) = \frac{Q(\theta = 0 \mid S = 1)}{Q(\theta = 1 \mid S = 1)} \times Q(S = 1, \theta = 1)$$

Thus, (2.5) is equivalent to

$$\frac{Q(S = 1, \theta = 0)}{Q(S = 1, \theta = 1)} = \frac{Q(\theta = 0 \mid S = 1)}{Q(\theta = 1 \mid S = 1)} \tag{2.6}$$

Again using Bayes' rule, the RHS can be rewritten

$$\frac{Q(\theta = 0 \mid S = 1)}{Q(\theta = 1 \mid S = 1)} = \frac{Q(\theta = 0, S = 1)/Q(S = 1)}{Q(\theta = 1, S = 1)/Q(S = 1)} = \frac{Q(S = 1, \theta = 0)}{Q(S = 1, \theta = 1)}$$

so (2.6) is equivalent to

$$\frac{Q(S = 1, \theta = 0)}{Q(S = 1, \theta = 1)} = \frac{Q(S = 1, \theta = 0)}{Q(S = 1, \theta = 1)}$$

and is therefore trivially true.

The identity (2.5) holds for both groups $g \in \{g_1, g_2\}$. So if $FP_{g_1} = FP_{g_2}$ (as required by equality of false positive rates), $FN_{g_1} = FN_{g_2}$ (as required by equality of false negative rates), and also $PPV_{g_1} = PPV_{g_2}$ (as required by calibration), it must also hold that $p_{g_1} = p_{g_2}$. ■

2.3 Gaussian Information

Gaussian information environments are unusually tractable, since the posterior belief can be expressed in closed-form. We'll cover the main formulae for Bayesian updating in these environments, and show how these can be used to derive results in three applications.

2.3.1 Formulae

We'll start with the simplest case. The state is $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ and the signal is $X = \theta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $\theta \perp\!\!\!\perp \varepsilon$, and $\sigma_\theta^2, \sigma_\varepsilon^2 > 0$. Then:

FACT 2.2. *The agent's posterior belief about θ conditional on signal realization $X = x$ is normally distributed with mean*

$$\mathbb{E}(\theta \mid X = x) = \left(\frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\right)\mu + \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\right)x$$

and variance

$$Var(\theta \mid X = x) = \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

A key property worth remembering is that the posterior mean is a convex combination of the prior mean μ and the signal realization x, where the weights are proportional to prior precision and signal precision. Additionally, while the posterior mean depends on the signal realization, the posterior variance is a constant.

Fact 2.2 is also sometimes written as:

$$(\theta \mid X = x) \sim \mathcal{N}\left(\left(\frac{\tau_\theta}{\tau_\theta + \tau_\varepsilon}\right)\mu + \left(\frac{\tau_\varepsilon}{\tau_\theta + \tau_\varepsilon}\right)x, \frac{1}{\tau_\theta + \tau_\varepsilon}\right)$$

where $\tau_\theta = 1/\sigma_\theta^2$ is the precision of the prior belief and $\tau_\varepsilon = 1/\sigma_\varepsilon^2$ is the precision of the signal. This restatement makes it apparent that the posterior precision is the sum of the prior precision and signal precision.

We can use Fact 2.2 to derive the distribution of the posterior mean.

EXERCISE 2.3 (U). *Let $\theta \sim \mathcal{N}(\mu, 1)$ and define two signals*

$$Y_1 = \theta + \varepsilon_1$$
$$Y_2 = \theta + \varepsilon_2$$

where θ, ε_1, and ε_2 are all independent of one another, and $\varepsilon_1 \sim \mathcal{N}(0, 1)$ while $\varepsilon_2 \sim \mathcal{N}(0, 2)$.

(a) Solve for the conditional distributions $\theta \mid Y_1 = y$ and $\theta \mid Y_2 = y$.

(b) What are the values of (μ, y) for which it the case that $\mathbb{E}(\theta \mid Y_1 = y) > \mathbb{E}(\theta \mid Y_2 = y)$? Provide intuition for the condition you derive.

EXERCISE 2.4 (U). *Suppose we write the posterior belief as $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu}$ is a random variable that depends on the realization of the signal X. Prove that*

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \sigma_\theta^2 - \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\right),$$

i.e. the expected posterior mean is the prior mean, and the variance of the posterior mean is equal to the prior variance (σ_θ^2), reduced by the posterior variance, $\left(\frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\right)$.

This characterization implies that the more informative the signal is, the more variable the posterior mean is.

REMARK 2.4. More generally (i.e., for θ and X that are not necessarily normally-distributed), the law of total variance says that

$$\mathrm{Var}(\mathbb{E}[\theta \mid X]) = \mathrm{Var}(\theta) - \mathbb{E}[\mathrm{Var}(\theta \mid X)]$$

so the variance of the posterior mean is equal to the difference of the prior variance and the expectation of the posterior variance.

Similar closed-forms exist for multivariate Gaussian states and signals. Suppose Z is a $1 \times K$ vector distributed according to $\mathcal{N}(\mu, \Sigma)$, where Σ has full rank. Partition the vector as follows:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

FACT 2.3. *The conditional distribution of Z_1 given $Z_2 = z_2$ is $\mathcal{N}(\hat{\mu}, \widehat{\Sigma})$ where*

$$\begin{aligned} \hat{\mu} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2) \\ \widehat{\Sigma} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Again, the posterior mean depends on the signal realization, but the posterior covariance matrix does not.

EXAMPLE 2.5. Let $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$. Then $(Z_1 \mid Z_2 = z_2) \sim \mathcal{N}(\hat{\mu}, \widehat{\Sigma})$ where

$$\begin{aligned} \hat{\mu} &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(z_2 - \mu_2) \\ \widehat{\Sigma} &= \sigma_1^2(1 - \rho^2) \end{aligned}$$

EXERCISE 2.5 (U). *Let $Z_1 = \theta$ and $Z_2 = X$ where θ and X are as defined at the beginning of this section. Show that Fact 2.3 implies Fact 2.2.*

Sections 2.3.2-2.3.4 demonstrate three applications of these Bayesian updating formulae.

2.3.2 Application 1: Career Concerns

Our first application is solving the two-period version of Holmstrom (1999) model of career concerns.

There is a single agent and a manager. The agent has a type $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ that is unknown to both the agent and the manager. In period 1, the agent chooses an effort level $a \in \mathbb{R}_+$ at cost $c(a) = \frac{1}{2}a^2$. This effort is not observed by the manager. The agent's type and effort jointly determine the realization of a performance signal

$$X = \theta + a + \varepsilon \tag{2.7}$$

where $\theta \perp\!\!\!\perp \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In period 2, the manager observes the realization of X and forms an expectation about the agent's type. Since the manager does not observe a, this expectation is taken with respect to the manager's possibly misspecified perception about the distribution of X (more soon). The agent receives the manager's expectation of his type.

For arbitrary $a \in \mathbb{R}_+$, write $\mathbb{E}^a(\theta \mid X)$ for the conditional expectation of θ with respect to $X = \theta + a + \varepsilon$. If the manager expects the agent to choose effort a^* while the agent in fact chooses effort a, then the agent's total expected payoff is

$$\mathbb{E}^a[\mathbb{E}^{a^*}(\theta \mid X)] - c(a),$$

where the inner expectation $\mathbb{E}^{a^*}(\theta \mid X)$ is the manager's expectation of the agent's type, and $\mathbb{E}^a[\mathbb{E}^{a^*}(\theta \mid X)]$ is the agent's expectation of the manager's expectation.

Claim 1. *There is a unique equilibrium in which the agent chooses effort* $a^* = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}$.

Corollary 2.2. *Equilibrium effort a^* is decreasing in σ_ε^2 (i.e., it is less valuable to manipulate a noisier signal) and is increasing in σ_θ^2 (i.e., it is more valuable to manipulate information about a more uncertain unknown).*

We'll now prove Claim 1. Equilibrium effort a^* must satisfy the first-order condition

$$\left.\frac{\partial \mathbb{E}^a[\mathbb{E}^{a^*}(\theta \mid X)]}{\partial a}\right|_{a=a^*} = a^* \tag{2.8}$$

equating the marginal value of increasing effort (over a^*) to the marginal cost of increasing effort (over a^*). Applying Fact 2.2, the manager's expectation of θ with respect to the de-biased signal $X - a^* = \theta + \varepsilon$ is

$$\mathbb{E}^{a^*}(\theta \mid X) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}(X - a^*) + \frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}\mu$$

The agent's expectation of this expectation (with respect to $X = \theta + a + \varepsilon$) is

$$\mathbb{E}^a\left[\mathbb{E}^{a^*}(\theta \mid X)\right] = \mu + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}(a - a^*)$$

So (2.8) implies that equilibrium effort is $a^* = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}$. (Uniqueness follows from strict concavity of the agent's payoff function.)

Exercise 2.6 (U). *Consider the model described in this section, and set* $\sigma_\theta^2 = \sigma_\varepsilon^2 = 1$.

(a) *Suppose that in addition to the worker's performance signal, the firm (through collection of additional data about the worker's type) is able to separately observe a signal*

$$S = \theta + \delta$$

where θ and δ are jointly normal and independent, and $\delta \sim \mathcal{N}(0,1)$. The firm's expectation about the worker's type is based both on S as well as on the worker's performance signal X (as defined in (2.7)). Solve for the worker's equilibrium action and compare it with the previous solution $a^* = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2}$. Does the worker exert more or less effort in equilibrium? Provide intuition for your result.

(b) *Suppose that the firm instead acquires data that allows it to more accurately monitor the performance shocks that the worker experiences (e.g., whether the worker had a rough day, or had help at work). Formally, the firm observes*

$$S = \varepsilon + \delta$$

where ε and δ are jointly normal and independent, and $\delta \sim \mathcal{N}(0,1)$. The firm's expectation about the worker's type is based both on S as well as on the worker's performance signal X (as defined in (2.7)). Solve for the worker's equilibrium action and compare it with the previous solution $a^* = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2}$. Does the worker exert more or less effort in equilibrium? Provide intuition for your result.

EXERCISE 2.7 (G). *Consider a variation on Holmstrom (1999)'s career concerns model, in which the type θ and noise term ε are correlated. Specifically, the type is decomposed as $\theta = \theta_1 + \theta_2$, the signal is $X = \theta + \varepsilon + a$, and we suppose that*

$$\theta_2 = \alpha\theta_1 + z$$
$$\varepsilon = \beta\theta_1 + w$$

where $\alpha, \beta \in \mathbb{R}$ are known constants, and $\theta_1 \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, $z \sim \mathcal{N}(0, \sigma_z^2)$, and $w \sim \mathcal{N}(0, \sigma_w^2)$ are mutually independent and unknown to both the agent and the manager.

(a) *Solve for equilibrium effort. How does this compare to Claim 1 in the special case $\alpha = \beta = 0$?*

(b) *Suppose $\alpha, \beta > 0$. How does equilibrium effort change in the parameters α and β? Provide intuition.*

2.3.3 Application 2: Linear-Quadatic Coordination Games

Our second application is solving for equilibrium in a two-agent linear-quadratic coordination game (Morris and Shin, 2002).

Let $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ be an unknown state. Each agent $i = 1, 2$ receives a private signal about the state

$$X_i = \theta + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is independent of the state and across agents. Each agent chooses an action $a_i \in \mathbb{R}$ given their signal realization x_i. Agent i's payoff is

$$U_i(a_1, a_2) = -(1-\beta)(a_i - \theta)^2 - \beta(a_i - a_j)^2$$

where $\beta \in (0, 1)$ controls how much the agent cares about matching the state versus matching the other agent's action.

We'll solve for a symmetric linear Bayesian Nash equilibrium (a_1^*, a_2^*) in which each agent's strategy satisfies

$$a_i^*(x_i) = cx_i + \kappa \tag{2.9}$$

for some constants $c, \kappa \in \mathbb{R}$. Let's first conjecture that such an equilibrium exists. Given agent j's strategy $a_j(x_j) = cx_j + \kappa$, agent i's expected payoff (conditional on $X_i = x_i$) is

$$\mathbb{E}[-(1-\beta)(a_i - \theta)^2 - \beta(a_i - (cX_j + \kappa))^2 \mid X_i = x_i]$$

Taking a derivative with respect to a_i, agent i's best reply is

$$a_i^*(x_i) = (1-\beta)\mathbb{E}(\theta \mid X_i = x_i) + \beta(c\mathbb{E}(\theta \mid X_i = x_i) + \kappa).$$

Plugging in the expression for $\mathbb{E}(\theta \mid X_i = x_i)$ from Fact 2.2, and matching coefficients with (2.9), we have $c = \frac{\sigma_\theta^2(1-\beta)}{\sigma_\varepsilon^2 + \sigma_\theta^2(1-\beta)}$ and $\kappa = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\theta^2(1-\beta)}\mu$. Thus a symmetric linear equilibrium exists in which each agent i chooses

$$a_i^*(x_i) = \frac{\sigma_\theta^2(1-\beta)}{\sigma_\varepsilon^2 + \sigma_\theta^2(1-\beta)} x_i + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\theta^2(1-\beta)}\mu \tag{2.10}$$

Morris and Shin (2002) further show that this is the unique pure-strategy equilibrium.

Suppose we interpret the common prior $\mathcal{N}(\mu, \sigma_\theta^2)$ as informed by a public signal, where a more informative signal implies a smaller σ_θ^2. Then we see from (2.10) that the more informative the public signal is, the less weight agents place on their private signal.

2.3.4 Application 3: Data Sharing

Our final application is an example from Acemoglu et al. (2022) regarding why online platforms don't compensate users for the data that they give up.

There is a single platform and two agents $i = 1, 2$ with types distributed

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Each agent i privately observes the realization of a signal $X_i = \theta_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$ is independent across agents and independent of both types.

The platform chooses a payment p_i to offer to each agent i for sharing their data. After receiving these offers, each agent i chooses whether to share $(a_i = 1)$ or withhold $(a_i = 0)$ their signal realization. Write $X_{\mathbf{a}}$ for the signals shared under action profile $\mathbf{a} = (a_1, a_2)$. For example, if $\mathbf{a} = (1, 0)$, then $X_{\mathbf{a}} = X_1$, while if $\mathbf{a} = (1, 1)$, then $X_{\mathbf{a}} = (X_1, X_2)$.

Each agent i's payoff is determined by the platform's posterior uncertainty about his type, a privacy parameter $v \in \mathbb{R}_+$, and his payment via

$$u_i(\mathbf{a}, \mathbf{p}) = v \cdot \mathrm{Var}(\theta_i \mid X_{\mathbf{a}}) + p_i \cdot \mathbb{1}(a_i = 1).$$

The platform's payoff is $u_P(\mathbf{a}, \mathbf{p}) = -u_1(\mathbf{a}, \mathbf{p}) - u_2(\mathbf{a}, \mathbf{p})$. So the agents prefer for the platform to be more uncertain about their types, while the platform prefers to be less uncertain.

We'll show that when agent types are sufficiently correlated, i.e., ρ is large, then the platform can induce both agents to share their data at a lower total payment than what is required to induce exactly one agent to share.

Let's first solve for payment vectors (p_1, p_2) given which it is an equilibrium for both agents to share their signals. Suppose agent j chooses to share. Then if agent i does not share, the platform's belief about θ_i is updated only to X_j. Since

$$\begin{pmatrix} \theta_i \\ X_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 2 \end{pmatrix} \right)$$

the platform's posterior variance of θ_i is $1 - \rho^2/2$ (by Fact 2.3). So agent i's payoff is $v \cdot (1 - \rho^2/2)$. If agent i does share, then beliefs about θ_i are further updated to the signal X_i, and (by Fact 2.2) the platform's posterior variance of θ_i reduces to $\frac{2-\rho^2}{4-\rho^2}$. So agent i's payoff is $v \cdot \left(\frac{2-\rho^2}{4-\rho^2}\right) + p_i$. Thus, agent i's best reply to $a_j = 1$ is to share if and only if

$$p_i \geq v \cdot \left(\frac{(2-\rho^2)^2}{2(4-\rho^2)} \right)$$

and the action profile $(a_1, a_2) = (1, 1)$ is an equilibrium if the above display holds for both agents i. The minimum total payment is twice the right-hand-side, i.e., $v \cdot \left(\frac{(2-\rho^2)^2}{(4-\rho^2)} \right)$.

Let's now solve for payment vectors (p_1, p_2) given which it is an equilibrium for exactly one agent to share his data. Without loss, fix $a_2 = 0$. If agent 1 chooses $a_1 = 0$, then the platform's uncertainty about θ_1 is its prior uncertainty, 1, so agent 1's payoff is v. If agent 1 chooses $a_1 = 1$, then the platform's belief about θ_1 updates to the signal X_1. Applying Fact 2.2, the platform's posterior variance about θ_1 is $1/2$ and so agent 1's payoff is $v \cdot (1/2) + p_1$. Thus, $a_1 = 1$ is a best reply to $a_2 = 0$ if and only if $p_1 \geq v/2$. So the platform can induce (exactly) one agent to share if it offers one agent a payment of at least $v/2$ (which is accepted) and another a payment of no more than $v/2$ (which is rejected), at a total payment of $v/2$.

When $\rho^2 \geq \frac{7-\sqrt{17}}{4} \approx 0.71$, then $v \cdot \left(\frac{(2-\rho^2)^2}{(4-\rho^2)}\right) < v/2$, so the platform pays less to induce two users share than one. Intuitively, each agent's choice to share their data exerts a negative externality on other agent: When both users share, each of their signals is less valuable in view of the signal revealed by the other. Agents paid their marginal value thus receive lower compensation, and in a limiting version of this model with a growing number of agents, the amount of compensation needed to induce all agents to share vanishes to zero.

2.4 Additional Exercises

EXERCISE 2.8 (U). *Suppose $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$ and*

$$\begin{aligned} Y_1 &= \theta + b + \varepsilon_1 \\ Y_2 &= b + \varepsilon_2 \end{aligned}$$

where θ, b, ε_1, and ε_2 are all independent of one another, $b \sim \mathcal{N}(0, \sigma_b^2)$, $\varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$, and $\varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$. We can interpret Y_1 as a biased signal about θ and Y_2 as a signal about the size of the bias.

Your friend says: "The only value of Y_1 and Y_2 for learning about θ is to provide information about the size of b. Since $Y_1 - Y_2$ is an unbiased signal about b, it is equally valuable to learn the outcome of $Y_1 - Y_2$ as it is to learn the pair of signals (Y_1, Y_2)."

Show that your friend is wrong: The distribution of $\theta \mid Y_1, Y_2$ is different from the distribution of $\theta \mid Y_1 - Y_2$. Also provide an intuition explaining to your friend the error in their reasoning.

EXERCISE 2.9. *Consider the two-player game described in Section 2.3.4, but suppose that the types are distributed*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

As in Section 2.3.4, let a_1 and a_2 denote the actions of players 1 and 2, where an action of '0' means that the player does not share their data, while '1' means that they do.

(a) Suppose player 2 chooses $a_2 = 0$. What is player 1's payoff from choosing $a_1 = 0$ and what is player 1's payoff from choosing $a_1 = 1$? Provide a condition that characterizes when it is the case that player 1's best reply is to share data (i.e., $a_1 = 1$).

(b) Suppose player 2 chooses $a_2 = 1$. What is player 1's payoff from choosing $a_1 = 0$ and what is player 1's payoff from choosing $a_1 = 1$? Provide a condition that characterizes when it is the case that player 1's best reply is to share data (i.e., $a_1 = 1$).

(c) Suppose player 1 chooses $a_1 = 0$. What is player 2's payoff from choosing $a_2 = 0$ and what is player 2's payoff from choosing $a_2 = 1$? Provide a condition that

characterizes when it is the case that player 2's best reply is to share data (i.e., $a_2 = 1$).

(d) *Suppose player 1 chooses $a_1 = 1$. What is player 1's payoff from choosing $a_2 = 0$ and what is player 2's payoff from choosing $a_2 = 1$? Provide a condition that characterizes when it is the case that player 2's best reply is to share data (i.e., $a_2 = 1$).*

(e) *Suppose $v = 1$. For what set of values of (p_1, p_2) is it the case that:*

(i) $(a_1, a_2) = (1,1)$ *is a Nash equilibrium?*

(ii) $(a_1, a_2) = (1,0)$ *is a Nash equilibrium?*

(iii) $(a_1, a_2) = (0,1)$ *is a Nash equilibrium?*

(iv) $(a_1, a_2) = (0,0)$ *is a Nash equilibrium?*

(f) *Again let $v = 1$. What is the smallest total payment the firm must make to induce an equilibrium where both players share their data? (That is, what is the smallest sum $p_1 + p_2$ such that $(a_1, a_2) = (1,1)$ is an equilibrium given the payment profile (p_1, p_2)?) Comment on whether it is the case that one player receives the higher payment, and why this answer makes sense.*

(g) *Again let $v = 1$. Suppose there is a firm 1 and firm 2, where firm 1 only interacts with player 1, and firm 2 only interacts with player 2. What is the smallest amount p_1 that firm 1 must pay to induce player 1 to choose $a_1 = 1$? What is the smallest amount p_2 that firm 2 must pay to induce player 2 to choose $a_2 = 2$? Compare the sum of these values $p_1 + p_2$ to your answer from part (f).*

Exercise 2.10 (G). *Suppose θ is normally distributed. For each $i = 1, \dots, n$, let $X_i = \theta + \varepsilon_i$ where ε_i is independent of θ, the vector $(\varepsilon_1, \dots, \varepsilon_n)$ is jointly normal, and the signals $X_1, \dots, X_n$ are exchangeable. Define $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Prove that $\theta \mid X_1, \dots, X_n$ is identical in distribution to $\theta \mid \overline{X}$.*

Hint. Recall that $\mathbb{E}(\theta \mid X)$ minimizes $\mathbb{E}[(\hat{\theta} - \theta)^2]$ among all $\sigma(X)$-measurable random variables $\hat{\theta}$.

Exercise 2.11 (G). *Consider two processes of social learning about an unknown state $\theta \sim \mathcal{N}(0,1)$.*

Scenario 1: *At $t = 0$, a single agent privately observes the signal*

$$Y = \theta + \delta, \quad \delta \sim \mathcal{N}(0, 1/\tau)$$

where θ and δ are independent of one another, and the precision $\tau \in \mathbb{R}_+$ is a known constant. The agent chooses an action y and receives the payoff $-\mathbb{E}[(y - \theta)^2]$. At $t = 1$, each of n agents, indexed by i, privately observes a signal

$$X_i = \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1)$$

as well as the action y of the first agent. The error terms ε_i are independent across agents, and independent of θ and δ. Each agent i from this generation then takes an *action a_i to maximize the payoff $-\mathbb{E}[(a_i - \theta)^2]$. At $t = 2$, you arrive, observe the actions $(a_1, \ldots, a_n)$ of the preceding generation (but* not *the action of the first agent), and choose an action a^* with payoff $-\mathbb{E}[(a^* - \theta)^2]$.*

Scenario 2: *At $t = 1$, each of m agents observes a private signal*

$$Z_i = \theta + \eta_i, \quad \eta_i \sim \mathcal{N}(0, 1)$$

where the error terms η_i are independent across agents and of θ. Each agent i takes an action b_i with payoff $-\mathbb{E}[(b_i - \theta)^2]$. At $t = 1$, you arrive, observe the actions $(b_1, \ldots, b_m)$ of the preceding generation, and choose an action a^* with payoff $-\mathbb{E}[(a^* - \theta)^2]$.

Characterize the function $h(n)$ such that your expected payoff is higher in scenario 1 if and only if $m < h(n)$. As clearly as you can, write out an intuition for this result.

HINT. Use the fact given in Exercise 2.10.

Chapter 3

Properties of Information

Many economic settings involve an unknown type or quality θ and a signal X about θ, where both θ and X are ordered (i.e., there are "better" qualities θ and "higher" signal realizations X). In these settings, we might think that higher realizations of X are good news about θ—for example, that higher test scores suggest higher ability or that better reviews for a product suggest higher quality. These positive inferences are not in general justified, requiring assumptions on the joint distribution of (θ, X).

Section 3.1 presents three useful definitions of positive dependence between random variables, which are applied to our motivating problem (inference about θ from observation of a signal X) in Section 3.2. Section 3.3 presents an example of the kind of counterintuitive result that can obtain when these properties are not imposed on the informational environment.

3.1 Definitions

3.1.1 Monotone Likelihood Ratio Property

Consider two random variables Z and $\widetilde{Z}$ with distributions F and $\widetilde{F}$ that admit densities f and $\tilde{f}$. To simplify exposition, all densities mentioned in this chapter are assumed to be everywhere positive.

Definition 3.1. *The distribution F* likelihood-ratio dominates *the distribution $\widetilde{F}$ if*[1]

$$\frac{f(z)}{f(z')} \geq \frac{\widetilde{f}(z)}{\widetilde{f}(z')} \qquad \textit{for all } z > z'$$

Intuitively, moving up in the likelihood-ratio dominance order renders higher realizations of z more likely relative to lower realizations.

This definition is often specialized to conditional densities in the following way. Suppose θ and X are real-valued random vectors defined on the same

[1] The assumption that densities are everywhere strictly positive allows us to define the monotone likelihood ratio property in terms of likelihood ratios. More generally, we can consider a distribution F to likelihood-ratio dominate another distribution $\widetilde{F}$ if $f(z)\tilde{f}(z') \geq f(z')\tilde{f}(z)$ for all $z > z'$.

probability space with densities f_θ and f_X and conditional densities $f_{\theta|X}$ and $f_{X|\theta}$.

Definition 3.2. *The family of conditional densities* $\{f_{X|\theta}(\cdot \mid \theta)\}_{\theta\in\Theta}$ *have the* monotone likelihood ratio property *(MLRP) if for every* $x > x'$ *and* $\theta > \theta'$,

$$\frac{f_{X|\theta}(x \mid \theta)}{f_{X|\theta}(x' \mid \theta)} \geq \frac{f_{X|\theta}(x \mid \theta')}{f_{X|\theta}(x' \mid \theta')}. \tag{3.1}$$

If the inequality above holds strictly at every $x > x'$, *then we say that* $\{f_{X|\theta}(\cdot \mid \theta)\}$ *have the* strict *monotone likelihood ratio property.*

Remark 3.1. If $\{f_{X|\theta}(\cdot \mid \theta)\}$ satisfy MLRP, then $\{f_{\theta|X}(\cdot \mid X)\}$ also satisfy MLRP. To see this, observe that by Bayes' rule, (3.1) can be rewritten

$$\frac{f_{\theta|X}(\theta \mid x)f_X(x)}{f_\theta(\theta)}\frac{f_\theta(\theta)}{f_{\theta|X}(\theta \mid x')f_X(x')} \geq \frac{f_{\theta|X}(\theta' \mid x)f_X(x)}{f_\theta(\theta')}\frac{f_\theta(\theta')}{f_{\theta|X}(\theta' \mid x')f(x')}$$

which simplifies to the condition that $\{f_{\theta|X}(\cdot \mid X)\}$ have the monotone likelihood ratio property.

In the special case of an additive signal $X = \theta + \varepsilon$, where ε is independent of θ and has density f_ε,

$$\frac{f_{\theta|X}(\theta \mid x)}{f_{\theta|X}(\theta' \mid x)} = \frac{f_\varepsilon(x-\theta)}{f_\varepsilon(x-\theta')}$$

so the MLRP condition in (3.1) becomes

$$\frac{f_\varepsilon(x-\theta)}{f_\varepsilon(x'-\theta)} \geq \frac{f_\varepsilon(x-\theta')}{f_\varepsilon(x'-\theta')} \qquad \text{for every } x > x' \text{ and } \theta > \theta',$$

i.e., for every $\theta > \theta'$, the function $\frac{f_\varepsilon(x-\theta)}{f_\varepsilon(x-\theta')}$ is nondecreasing in x. It turns out that this is precisely the condition that f_ε is log concave.

Definition 3.3. *A function f that maps a convex set into the positive reals is* log-concave *if the function* $\ln f$ *is concave.*

Proposition 5 (Saumard and Wellner (2014)). *A density function f on $\mathbb{R}$ is log-concave if and only if for every $\theta > \theta'$, the ratio $\frac{f(x-\theta)}{f(x-\theta')}$ is a non-decreasing function of x.*

Thus, in any model where (1) $X = \theta + \varepsilon$, (2) the noise term ε is independent of θ, and (3) ε has a log-concave density, we can be guaranteed that $\{f_{\theta|X}(\cdot \mid x)\}$ has the monotone likelihood ratio property (no matter the distribution of θ).

Many distributions have log-concave densities—for example, normal distributions, exponential distributions, the uniform distribution over any convex set, the logistic distribution, and the extreme value distribution. But others do not—for example, the Pareto distribution and Cauchy distribution. See Saumard and Wellner (2014) or Bagnoli and Bergstrom (2005) for other examples and properties of log-concave distributions.

3.1.2 Affiliation

Let $Z_1, \dots, Z_n$ be real-valued random variables taking values in $\mathbb{R}^n$ and admitting joint density f, which again we'll assume to be everywhere strictly positive. For any $z, z' \in \mathbb{R}^n$, let $z \wedge z'$ ("z meet z′) denote the component-wise minimum of z and z', and $z \vee z'$ ("z join z′) denote the component-wise maximum, i.e.,

$$\begin{aligned} z \vee z' &= (\max(z_1, z_1'), \dots, \max(z_n, z_n')) \\ z \wedge z' &= (\min(z_1, z_1'), \dots, \min(z_n, z_n')) \end{aligned}$$

Definition 3.4. *The variables $Z_1, \dots, Z_n$ are* affiliated *if*

$$f(z \vee z') f(z \wedge z') \geq f(z) f(z') \tag{3.2}$$

for all $z, z' \in \mathbb{R}^n$.

This condition loosely says that larger realizations of any one variable make larger realizations of the other variables more likely. Figure 3.1 depicts this relationship for two binary variables.

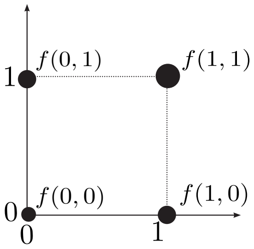

Figure 3.1: Two binary variables with joint density f are affiliated if $f(1,1)f(0,0) \geq f(1,0)f(0,1)$.

Remark 3.2. If $Z_1, \dots, Z_n$ are mutually independent, then they are affiliated.

Besides Definition 3.4, there are several equivalent ways to characterize affiliation. The first follows by taking logs of both sides of (3.2).

Proposition 6. *$Z_1, Z_2, \dots, Z_n$ are affiliated if and only if f is log-supermodular, i.e.*

$$\log f(z \vee z') + \log f(z \wedge z') \geq \log f(z) + \log f(z')$$

for all z, z'.

Proposition 7. *Suppose the joint density f is twice-differentiable. Then $Z_1, Z_2, \dots, Z_n$ are affiliated if and only if $\frac{\partial^2 \log f}{\partial z_i z_j} \geq 0$.*

We show the only if direction below, leaving the if direction for an exercise.

Proof. Without loss let $i = 1$ and $j = 2$. Choose any $z_1, z_1', z_2, z_2' \in \mathbb{R}$ where $z_1 > z_1'$ and $z_2 > z_2'$. Suppose $Z_1, Z_2, \dots, Z_n$ are affiliated. Then by definition

$$\log f(z_1, z_2, z_{-12}) - \log f(z_1', z_2, z_{-12}) \geq \log f(z_1, z_2', z_{-12}) - \log f(z_1', z_2', z_{-12})$$

where z_{-12} is shorthand for $(z_3, \dots, z_n)$. Rewrite z_1 as $z_1' + \varepsilon$ and divide both sides by ε. Taking the limit as $\varepsilon \to 0$, we have

$$\lim_{\varepsilon \to 0} \left(\frac{\log f(z_1' + \varepsilon, z_2, z_{-12}) - \log f(z_1', z_2, z_{-12})}{\varepsilon} \right) \geq \lim_{\varepsilon \to 0} \left(\frac{\log f(z_1' + \varepsilon, z_2', z_{-12}) - \log f(z_1', z_2', z_{-12})}{\varepsilon} \right)$$

so $\frac{\partial \log f}{\partial z_1}$ is increasing in z_2, as desired. ■

Exercise 3.1 (G). *Prove the 'if' direction of Proposition 7: If the joint density f is twice-differentiable and satisfies $\frac{\partial^2 \log f}{\partial z_i z_j} \geq 0$, then $Z_1, Z_2, \dots, Z_n$ are affiliated.*

The next characterization simplifies (3.2) to a pairwise condition. Specifically, for any (Z_i, Z_j) and any realization of the remaining variables Z_{-ij}, higher realizations of Z_i must imply higher realizations of Z_j.

Proposition 8. *$Z_1, \dots, Z_n$ are affiliated if and only if*

$$f(z_i, z_j, z_{-ij}) f(z_i', z_j', z_{-ij}) \geq f(z_i', z_j, z_{-ij}) f(z_i, z_j', z_{-ij}) \tag{3.3}$$

for every pair of distinct indices i, j, and every $z_i > z_i'$, $z_j > z_j'$, and $z_{-ij} \in \mathbb{R}^{n-2}$.

Exercise 3.2 (G). *Prove Proposition 8.*

This pairwise characterization immediately implies the following characterization, which says that $(Z_1, \dots, Z_n)$ are affiliated if and only if for every pair of variables i, j, and every realization of z_{-ij}, the family of conditional densities $\{f(\cdot \mid z_j, z_{-ij})\}_{z_j \in \mathbb{R}}$ has the monotone-likelihood ratio property.

Proposition 9. *$Z_1, \dots, Z_n$ are affiliated if and only if*

$$f(z_i \mid z_j, z_{-ij}) f(z_i' \mid z_j', z_{-ij}) \geq f(z_i \mid z_j', z_{-ij}) f(z_i' \mid z_j, z_{-ij}) \tag{3.4}$$

for every pair of distinct indices i, j, and every $z_i > z_i'$, $z_j > z_j'$, and $z_{-ij} \in \mathbb{R}^{n-2}$.

Proof. The displays in (3.3) and (3.4) are equivalent to one another by Bayes' rule, so Proposition 8 implies Proposition 9. ■

Operations that preserve affiliation include:

Proposition 10 (Monotone Functions). *Suppose $Z_1, \dots, Z_n$ are affiliated, and the functions $g_i : \mathbb{R} \to \mathbb{R}$, $1 \leq i \leq n$, are either all nondecreasing or all nonincreasing. Then the variables $g_1(Z_1), \dots, g_n(Z_n)$ are affiliated.*

Proposition 11 (Subsets). *Suppose $Z_1, \dots, Z_n$ are affiliated and let $A \subseteq \{1, \dots, n\}$ be any subset of these variables. Then the variables $(Z_i)_{i \in A}$ are affiliated.*

Proposition 12 (Order Statistics). *For each $1 \leq i \leq n$, let $z^{(i)}$ denote the i-th largest realization among $(z_1, \dots, z_n)$. Then the variables $(Z^{(1)}, \dots, Z^{(n)})$ are affiliated.*

Exercise 3.3 (G). *Show that affiliation is not preserved under arbitrary linear combinations of affiliated variables by constructing an example of random variables Z_1, Z_2, Z_3 where (Z_1, Z_2, Z_3) are affiliated but $(Z_1 + Z_2, Z_3)$ are not.*

3.1.3 First-Order Stochastic Dominance

Again consider two real-valued random variables, a parameter θ and a signal X, defined on the same probability space with joint distribution F. In many applications we may expect a higher signal realization to lead to a higher inference about the unknown parameter. We now formalize 'higher inference' as a first-order stochastic dominance shift in the posterior belief.

Definition 3.5. *A distribution F* first-order stochastically dominates $\widetilde{F}$, *which we denote by $F \geq_{FOSD} \widetilde{F}$, if*

$$\int u(\theta)dF(\theta) \geq \int u(\theta)d\widetilde{F}(\theta)$$

for every nondecreasing function $u : \mathbb{R} \to \mathbb{R}$. Equivalently, $F(\theta) \leq \tilde{F}(\theta)$ at every $\theta \in \Theta$.

If u is interpreted as a utility function over money, then a monetary gamble distributed according to F is preferred over one distributed according to $\tilde{F}$ by every agent that prefers more money over less, regardless of the specific shape of the agent's utility function. We can use this definition to compare conditional beliefs about θ.

Definition 3.6. *Say that F has the* FOSD property *if $F_{\theta|X}(\cdot \mid X = x) \geq_{FOSD} F_{\theta|X}(\cdot \mid X = x')$ for all $x > x'$.*

Milgrom (1981) proposed a closely related property, which is imposed on conditional distributions $F_{X|\theta}$ rather than joint distributions F. (This is analogous to considering a signal $\sigma : \Theta \to \Delta(S)$ without fixing a prior on Θ.)

Definition 3.7. *Say that a signal realization x is* more favorable than *signal realization x' if for every prior distribution $F_\theta \in \Delta(\Theta)$, the posterior distribution $F_{\theta|X}(\cdot \mid x)$ first-order stochastically dominates the posterior distribution $F_{\theta|X}(\cdot \mid x')$.*

That is, x is more favorable than x' if observing the realization x leads to a FOSD-higher posterior belief about θ (compared to observing x'). If x is more favorable than x' for all $x > x'$, then we have a stronger version of the FOSD property (given in (3.6)) that holds not only for the specific joint distribution F, but for all joint distributions F that are generated by $F_{X|\theta}$ and some choice of prior F_θ.

Example 3.1. Recall that in the normal-updating setting with $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$, $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and $\theta \perp\!\!\!\perp \varepsilon$, the agent's posterior belief about θ conditional on X is

$$\mathcal{N}\left(\frac{\sigma_\theta^2}{\sigma_\varepsilon^2 + \sigma_\theta^2}X + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\theta^2}\mu, \frac{\sigma_\varepsilon^2\sigma_\theta^2}{\sigma_\varepsilon^2 + \sigma_\theta^2}\right).$$

This distribution is increasing (in the FOSD order) in the realization of X for all parameters μ and σ_θ^2. So x is more favorable than x' for every pair $x > x'$.

3.2 How They are Related

Let θ and X be real-valued random vectors defined on the same probability space. We'll use F to denote their joint distribution, and assume throughout that the densities f_θ and f_X and conditional densities $f_{\theta|X}$ and $f_{X|\theta}$ exist. In this setting, our main properties from above are:

A: (X,θ) are affiliated.

MLRP: $\{f_{X|\theta}(\cdot \mid \theta)\}$ satisfies MLRP.

FOSD: For all $x > x'$, $F(\cdot \mid X = x) \geq_{FOSD} F(\cdot \mid X = x')$

MF: For all $x > x'$, x is more favorable than x'

These properties are related in the following way:

$$\textbf{(A)} \iff \textbf{(MLRP)} \iff \textbf{(MF)} \implies \textbf{(FOSD)}$$

where the one-directional implication from (MF) to (FOSD) is strict. See de Castro (2009) for an example of a distribution satisfying (FOSD) but not (MLRP).

Remark 3.3. (MLRP) is equivalent to (MF) but strictly stronger than (FOSD). Thus if a joint distribution F satisfies (MLRP) then it must satisfy (FOSD), but F can satisfy (FOSD) and fail (MLRP). On the other hand, a conditional distribution $F_{X|\theta}$ that satisfies (FOSD) for every completion to a joint distribution F (i.e., for every choice of prior F_θ) must also satisfy (MLRP). So ″FOSD for every prior″ is equivalent to MLRP, while "FOSD for some prior" is weaker.

We've already established the equivalence between (A) and (MLRP) in Proposition 9. Since (FOSD) is necessary for (MF), clearly (MF) implies (FOSD). The following result proves equivalence of (MLRP) and (MF).

Proposition 13 (Milgrom (1981))**.** *x is more favorable than x' if and only if for every $\theta > \theta'$,*

$$\frac{f_{X|\theta}(x \mid \theta)}{f_{X|\theta}(x' \mid \theta)} \geq \frac{f_{X|\theta'}(x \mid \theta')}{f_{X|\theta'}(x' \mid \theta')} \tag{3.5}$$

Proof. We will first show that if (3.5) is satisfied at every $\theta > \theta'$, then x must be more favorable than x'. Fix any prior F_θ and parameter $\theta^* \in \Theta$. If $F_\theta(\theta^*) \in \{0,1\}$ then the conclusion is trivially reached. So suppose $F_\theta(\theta^*) \in (0,1)$.

For any $\theta \leq \theta^*$ and $\tilde{\theta} > \theta^*$, (3.5) implies

$$\frac{f(x \mid \tilde{\theta})}{f(x \mid \theta)} \geq \frac{f(x' \mid \tilde{\theta})}{f(x' \mid \theta)}$$

where we omit subscripts on the densities here and elsewhere in the proof to ease notation. Integrating over all $\tilde{\theta}$ such that $\tilde{\theta} > \theta^*$ (with respect to the prior distribution F_θ), we obtain

$$\frac{\int_{\tilde{\theta}>\theta^*} f(x \mid \tilde{\theta}) dF_\theta(\tilde{\theta})}{f(x \mid \theta)} \geq \frac{\int_{\tilde{\theta}>\theta^*} f(x' \mid \tilde{\theta}) dF_\theta(\tilde{\theta})}{f(x' \mid \theta)}$$

or equivalently

$$\frac{f(x \mid \theta)}{\int_{\tilde{\theta}>\theta^*} f(x \mid \tilde{\theta})dF_\theta(\tilde{\theta})} \leq \frac{f(x' \mid \theta)}{\int_{\tilde{\theta}>\theta^*} f(x' \mid \tilde{\theta})dF_\theta(\tilde{\theta})}.$$

Integrating over all θ such that $\theta \leq \theta^*$, we obtain

$$\frac{\int_{\theta\leq\theta^*} f(x \mid \theta)dF_\theta(\theta)}{\int_{\tilde{\theta}>\theta^*} f(x \mid \tilde{\theta})dF_\theta(\tilde{\theta})} \leq \frac{\int_{\theta\leq\theta^*} f(x' \mid \theta)dF_\theta(\theta)}{\int_{\tilde{\theta}>\theta^*} f(x' \mid \tilde{\theta})dF_\theta(\tilde{\theta})}.$$

Recall that $f(x \mid \theta)f(\theta) = f(\theta \mid x)f(x)$, so the above display implies

$$\frac{\int_{\theta\leq\theta^*} f(\theta \mid x)d\theta}{\int_{\tilde{\theta}>\theta^*} f(\tilde{\theta} \mid x)d\tilde{\theta}} \leq \frac{\int_{\theta\leq\theta^*} f(\theta \mid x')d\theta}{\int_{\tilde{\theta}>\theta^*} f(\tilde{\theta} \mid x')d\tilde{\theta}}$$

or more simply

$$\frac{F(\theta^* \mid x)}{1 - F(\theta^* \mid x)} \leq \frac{F(\theta^* \mid x')}{1 - F(\theta^* \mid x')}$$

Since $\frac{y}{1-y}$ is a strictly increasing function in y, we have $F(\theta^* \mid x) \leq F(\theta^* \mid x')$ as desired.

In the other direction, we will show that if x is more favorable than x', then (3.5) holds everywere. Consider any two parameter values $\theta > \theta'$, and let F_θ be a prior distribution supported on these two points with equal probability on each.

Since by assumption x is more favorable than x', then $F(\theta' \mid x) \leq F(\theta' \mid x')$, implying

$$\frac{F(\theta' \mid x)}{1 - F(\theta' \mid x)} \leq \frac{F(\theta' \mid x')}{1 - F(\theta' \mid x')}$$

or equivalently

$$\frac{f(\theta' \mid x')}{f(\theta \mid x')} \geq \frac{f(\theta' \mid x)}{f(\theta \mid x)}.$$

Applying Bayes' rule again, we can rewrite the above as $\frac{f(x|\theta)}{f(x'|\theta)} \geq \frac{f(x|\theta')}{f(x'|\theta')}$, which is the desired conclusion. ■

Remark 3.4. Milgrom (1981)'s result is not precisely the proposition above, but instead the equivalence between strict MLRP (as defined in Definition 3.2) and a definition of "more favorable" that replaces FOSD with strict FOSD.

Specifically, say that F strictly first-order stochastically dominates $\widetilde{F}$ if $F(\theta) \leq \widetilde{F}(\theta)$ everywhere with strict inequality at some θ. (Equivalently, $\int u(\theta)dF(\theta) > \int u(\theta)d\widetilde{F}(\theta)$ for every strictly increasing function $u : \mathbb{R} \to \mathbb{R}$.) Say that x is strictly more favorable than x' if for every prior distribution F_θ, the posterior distribution $F_{\theta|X}(\cdot \mid x)$ strictly first-order stochastically dominates $F_{\theta|X}(\cdot \mid x')$. Then, by substituting strict inequalities in place of weak inequalities in the proof above where appropriate, we can conclude that $\{f_{X|\theta}(\cdot \mid \theta)\}$ satisfies strict MLRP if and only if x is strictly more favorable than x'.[2]

[2] Indeed, the same proof demonstrates a stronger (if slightly more cumbersome to state) result: If and only if $\{f_{X|\theta}(\cdot \mid \theta)\}$ satisfies strict MLRP, then $F_{\theta|X}(\theta \mid x) < F_{\theta|X}(\theta \mid x')$ at every θ such that $0 < F(\theta) < 1$.

We conclude by briefly summarizing other notions of positive dependence and placing the above properties relative to these.

Positive covariance (C): $Cov(X, \theta) \geq 0$

Positive quadrant dependence (QD): $Cov(g(X), h(\theta)) \geq 0$ for all non-decreasing functions g and h

Association (As): $Cov(g(X, \theta), h(X, \theta)) \geq 0$ for all non-decreasing functions g and h

Left-Tail Decreasing (LT): For all x, $F_{X|\theta}(X \leq x \mid \theta \leq t)$ is non-increasing in t, and for all t, $F_{\theta|X}(\theta \leq t \mid X \leq x)$ is non-increasing in x.

Inverse Hazard Rate Decreasing (IHR): For all x, $F_{X|\theta}(x \mid t)/f_{X|\theta}(x \mid t)$ is non-increasing in t, and for all t, $F_{\theta|X}(t \mid x)/f_{\theta|X}(t \mid x)$ is non-increasing in x.

These properties are extensively studied in, for example, Lehmann (1966), Esary, Proschan and Walkup (1967), de Castro (2009), and Chapter 3 of Balakrishna and Lai (2009). The following chain of implications is summarized in de Castro (2009):

Theorem 3.1. (A) $\iff$ (MLRP) $\implies$ (IHR) $\implies$ (FOSD) $\implies$ (LT) $\implies$ (As) $\implies$ (QD) $\implies$ (C)

Thus the standard properties of affiliation and MLRP are in fact strong, implying all of the other properties but not in general implied by them. These properties are equivalent to one another in the special case in which the two variables are jointly normal.

Exercise 3.4 (G). *Suppose $(X_1, \ldots, X_n)$ are jointly normal and exchangeable, where $\sigma^2 = Var(X_i)$ for each i, and $\rho = Cov(X_i, X_j)$ for each pair of indices i, j. Prove that these variables are affiliated if and only if $\rho \geq 0$.*

Hint. Use the fact given in Exercise 2.10.

3.3 When These Conditions Fail

An example from Lagziel and Lehrer (2019) demonstrates the kind of counter-intuitive result that can hold in settings where (A) and (MLRP) fail.

An editor chooses which papers to publish. Papers have unknown quality graded on a 9-point scale (A+, A, A-, B+, B, B-, C+,C,C-), whose prior distribution is given in Figure 3.2.

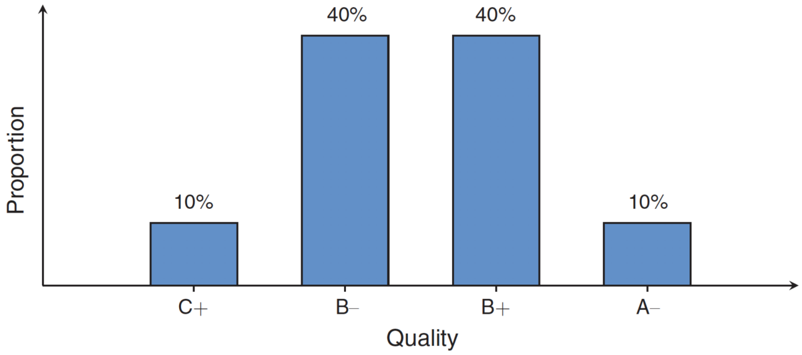

Figure 3.2: Distribution of Papers' Quality

The editor learns about quality via a noisy refereeing process, which generates an unbiased signal X about the paper. The realization of X is equal to the true quality with probability 0.8, and otherwise exactly two levels higher or lower than the true quality (each with probability 0.1). The distribution of X is reported in Figure 3.3:

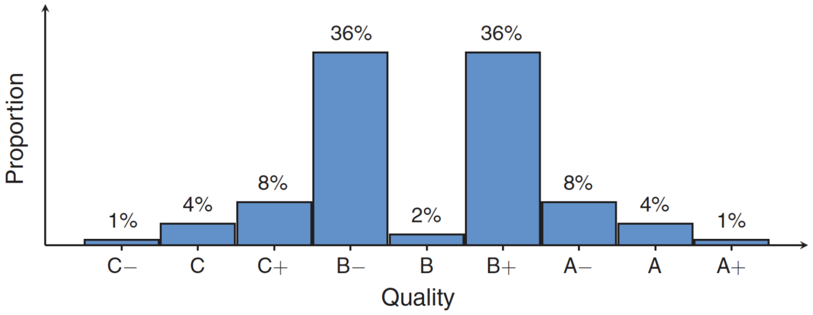

Figure 3.3: Distribution of Referee Signal

The editor chooses a threshold and accepts all papers whose expected quality (given the referee's report) exceeds this threshold. Intuitively, we may expect that the editor faces a tradeoff between publishing more papers versus publishing higher quality papers, where a higher threshold corresponds to publishing fewer but higher quality papers.

But observe that if the editor chooses to publish only papers with an expected quality that (weakly) exceeds A (i.e., the top-rated 5% of papers), then the expected value of the published work is close to $B+$. If the editor lowers the bar to $A-$ (i.e., the top-rated 13%), then the expected value of the published work *increases* to $A-$. Not only are more papers published, but their expected quality is higher.

In this example, we have $\mathbb{E}(\theta \mid X = x) < \mathbb{E}(\theta \mid X = x')$ even while $x > x'$, so clearly the posterior belief at x' does not first-order stochastically dominate the posterior belief at x. Chambers and Healy (2011) demonstrate an even stronger reversal by constructing signals such that the posterior belief at the lower signal realization first-order stochastically dominates the posterior belief at the higher signal realization. Notably, their result relies on natural-seeming signals that satisfy various reasonable properties.

Theorem 3.2. *For every non-degenerate, bounded θ there exists a signal structure X and two signal realizations $x' > x$ such that $f(\theta \mid X = x')$ is strictly first-order stochastically dominated by $f(\theta \mid X = x)$. Furthermore, X can be chosen to have the following properties: i) X is an additive signal structure, and ii) $e := X - \theta$ is mean-zero, symmetric, quasiconcave, and has bounded support.*

See Heinsalu (2020) for a strengthening of Lagziel and Lehrer (2019)'s example using this result, in which lowering the threshold not only increases the expected quality, but results in a quality distribution for published papers that first-order stochastically dominates the one that would obtain at the higher threshold.

3.4 Additional Exercises

Exercise 3.5 (G). *Let $Z_1, \ldots, Z_n$ be affiliated and let $h : \mathbb{R}^n \to \mathbb{R}$ be any function that is nondecreasing in each of its coordinates. Prove that the function*

$$\mathbb{E}(h(Z_1, \ldots, Z_n) \mid Z_1 = z_1)$$

is nondecreasing in z_1.

Exercise 3.6 (G). *Let X be any real-valued random variable and let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be bounded nondecreasing functions. Prove that $Cov(f(X), g(X)) \geq 0$. (Do not apply the FKG inequality.)*

Hint. There are at least two short proofs, one that uses Fubini's theorem and the fact that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ for any independent random variables X and Y, and another which relies entirely on elementary (if not obvious) arguments.

Exercise 3.7 (G). *Ann and Bob share the same prior p over an unknown real-valued state θ, and observe a common realization of the signal X, but disagree about the distribution of S. Ann believes that $X = \theta + \varepsilon$, where $\theta \perp\!\!\!\perp \varepsilon$ and ε is a real-valued noise term with density f_ε. Bob believes that $X = \theta + \varepsilon + \Delta$ for some $\Delta > 0$. That is, Ann perceives Bob as adding Δ to the realization of the signal, while Bob perceives Ann as subtracting Δ from the realization of the signal.*

Let f_A denote the joint density of (θ, X) according to Ann's model and f_B denote the joint density according to Bob's model, with $\mathbb{E}^A$ and $\mathbb{E}^B$ denoting their respective expectation operators. Impose the monotone likelihood-ratio property on $\{f_A(\cdot \mid \omega)\}$, that is,

$$\frac{f_A(x' \mid \theta')}{f_A(x \mid \theta')} \geq \frac{f_A(x' \mid \theta)}{f_A(x \mid \theta)} \quad \forall x' > x, \theta' > \theta$$

(a) Prove that $\{f_B(\cdot \mid \theta)\}$ also satisfies MLRP.

(b) Prove that $\mathbb{E}^A[\mathbb{E}^B[\theta \mid X]]$ is decreasing in Δ, and interpret this result.

(c) *Suppose that Ann and Bob now additionally observe a common vector of iid signals* $(Y_1, Y_2, \ldots, Y_N)$ *where each* $Y_i = \theta + \delta_i$ *with* $\theta \perp\!\!\!\perp \delta_i$ *and* δ_i *are iid across signals. Prove that*

$$\mathbb{E}^A[\mathbb{E}^B[\theta \mid X, Y_1, \ldots, Y_N]] \leq \mathbb{E}^A[\mathbb{E}^B[\theta \mid X, Y_1, \ldots, Y_N, Y_{N+1}]]$$

for every $N \geq 1$. *Again, interpret the result.*

Chapter 4

Comparing Information I: The Blackwell Order

When an agent has access to a choice between multiple signals, we may desire to order these signals based on how informative they are. Intuition can guide us on how to define such an ordering in specific cases, for example:

- Adding noise to a signal decreases its informativeness.
- Observing the realization of (X, Y) is more informative than observing the realization of X alone.

Any informativeness ordering should satisfy these properties, but there are different ways to generalize from here. One approach is to fix a decision problem and characterize the instrumental value of the signal for that decision problem. Alternatively, we could look for a universal informativeness ordering over signals that holds for all decision problems (as will be the focus of this chapter). Yet another approach is to quantify the "signal content" contained within the signal based on the physical difficulty of producing or processing that information (see Chapter 5).

In this chapter we introduce the Blackwell partial order on signals, which considers one signal to be more informative than another if it is more useful for all decision problems. If σ dominates σ' in this Blackwell order, we will say that *σ is more informative than σ'* or that *σ Blackwell-dominates σ'*. The following sections demonstrate five perspectives on this order, culminating in Blackwell's theorem (establishing their equivalence) and the proof of this theorem.

4.1 Garblings

We may consider a signal to be more informative than another if the latter is a noised-up version of the former.

Definition 4.1 (Markov matrix). *A matrix M is a* Markov matrix *if its entries are nonnegative and its rows sum to 1.*

Recall that when the set of states and the set of signal realizations are finite, we can represent any signal as a Markov matrix.

Definition 4.2 (Garblings, Finite Version). *Markov matrix P is a* garbling *of Markov matrix Q if there exists a Markov matrix M s.t. $QM = P$*

Example 4.1. Let $\Theta = \{\theta_1, \theta_2\}$ and consider the signals

$$P = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \qquad Q = \begin{pmatrix} 9/16 & 3/16 & 3/16 & 1/16 \\ 1/16 & 3/16 & 3/16 & 9/16 \end{pmatrix}$$

where as usual the rows are indexed to states and the columns are indexed to signal realizations. Then since

$$\underbrace{\begin{pmatrix} 9/16 & 3/16 & 3/16 & 1/16 \\ 1/16 & 3/16 & 3/16 & 9/16 \end{pmatrix}}_{Q} \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}_{M} = \underbrace{\begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}}_{P}$$

where M is a Markov matrix, we can conclude that P is a garbling of Q.

This example has a particularly nice intuition. Label the possible realizations of the first information structure P as s_1 and s_2, and consider the signal which is two independent realizations of P. The set of possible realizations of this new signal is then $\{s_1s_1, s_1s_2, s_2s_1, s_2s_2\}$ with the conditional distributions over these realizations given precisely by Q. So observing P is statistically equivalent to observing Q and forgetting the second realization. Clearly then P is less informative than Q.

Exercise 4.1 (U). *The state space is $\Theta = \{\theta_1, \theta_2\}$ and two signals are described by the following signal structures*

$$P = \left(\begin{array}{cccc} & s_1 & s_2 & s_3 \\ \theta_1 & 2/3 & 1/3 & 0 \\ \theta_2 & 0 & 1/3 & 2/3 \end{array}\right)$$

$$Q = \left(\begin{array}{ccc} & \tilde{s}_1 & \tilde{s}_2 \\ \theta_1 & 1/6 & 5/6 \\ \theta_2 & 5/6 & 1/6 \end{array}\right)$$

Show that Q is a garbling of P. (How does this exercise relate to Example 4.1?)

Exercise 4.2 (G). *For any $q \in [1/2, 1]$, define the matrix*

$$S_q = \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}$$

Suppose $q, q' \in [1/2, 1]$ with $q > q'$. Prove that $S_{q'}$ is a garbling of $S_{q'}$.

More generally, we can replace the Markov matrix M in Definition 4.2 with a Markov kernel.

Definition 4.3 (Garblings, General Version). *The signal $\sigma' : \Theta \to \Delta(S')$ is a* garbling *of the signal $\sigma : \Theta \to \Delta(S)$ if there exists a Markov kernel $\gamma : S \to \Delta(S')$ such that*

$$\sigma'(s' \mid \theta) = \int_{s \in S} \gamma(s' \mid s) \sigma(s \mid \theta) ds$$

Example 4.2. Let θ, ε, and δ be independent real-valued random variables with densities f_θ, f_ε, and f_δ. Then the signal $X = \theta + \varepsilon + \delta$ is a garbling of $Y = \theta + \varepsilon$, since

$$f_{X|\theta}(x \mid t) = \int_{y \in \mathbb{R}} f_\delta(x - y) f_{Y|\theta}(y \mid t) dy$$

where f_δ is a Markov kernel.

Example 4.3. Consider an arbitrary finite set Θ and let I be the $|\Theta| \times |\Theta|$ identity matrix. Then for any set of signal realizations S and any $|\Theta| \times |S|$ Markov matrix Q, we have $IQ = Q$, so Q is a garbling of I.

Exercise 4.3 (U). *Is it possible for P and Q to both be garblings of one another if $P \neq Q$? Provide an example if so, and otherwise prove that it is not possible.*

Remark 4.1. Let X and X' respectively denote the random realizations of the signals σ and σ'. Then θ, X, and X' are random variables which can be defined on a common probability space. The property that σ' is a garbling of σ does not however pin down the joint distribution of (θ, X, X'). What it guarantees is that there is a way of generating these variables such that θ is independent of X' conditional on X, in which case $\theta \mid X$ is identical in distribution to $\theta \mid X, X'$.[1] Other ways of generating these variables—still consistent with σ' being a garbling of σ—can yield different relationships.

For example, suppose $\theta \sim \mathcal{N}(0, 1)$ while

$$X = \theta + \varepsilon_1$$
$$X' = \theta + \varepsilon_2$$

where $\varepsilon_1 \sim \mathcal{N}(0, 1)$ and $\varepsilon_2 \sim \mathcal{N}(0, 2)$ are both independent of θ. Then clearly the latter signal is a garbling of the former. If we further assume that $\varepsilon_2 = \varepsilon_1 + \delta$ where $\delta \sim \mathcal{N}(0, 1)$ is an independent noise term, then the following statements are true:

[1] First draw the state θ, then draw X according to its conditional distribution, and finally draw X' according to the garbling kernel γ, independent of θ.

- X' is independent of θ conditional on X.
- X' is not independent of X conditional on θ (since they are further related through the common component ε_1).

On the other hand, if we assume that ε_1 and ε_2 are independent, then the statements above are reversed:

- X' is not independent of θ conditional on X (since X' provides additional information about θ beyond what is revealed by X).
- X' is independent of X conditional on θ.

Thus in general, the assumption that two signals are related by a garbling does not imply either conditional independence statement given above.

4.2 Decision Problems

Our next two definitions are based on the instrumental value of the signal for decision problems.

Definition 4.4. *A decision problem is any pair* $\mathbf{D} = (A, u)$ *where* A *is an action set and* $u : A \times \Theta \to \mathbb{R}$ *is a payoff function.*

The full decision problem is described as follows. Fix a prior $p \in \Delta(\Theta)$ and a signal $\sigma : \Theta \to \Delta(S)$.

1. The agent chooses a strategy $\alpha : S \to A$.
2. The state $\theta \sim p$ and signal realization $s \sim \sigma(\cdot \mid \theta)$ are realized, and the agent takes action $\alpha(s)$. The agent's payoff is $u(\alpha(s), \theta)$.

Without the benefit of further information, the best expected payoff the agent can achieve is

$$\sup_{a \in A} \mathbb{E}\left[u(a, \theta)\right] \tag{4.1}$$

With the benefit of the signal, the agent can achieve an expected payoff of

$$\sup_{\alpha : S \to A} \mathbb{E}\left[u(\alpha(s), \theta)\right] = \mathbb{E}\left[\sup_{a \in A} \mathbb{E}\left[u(a, \theta) \mid s\right]\right] \tag{4.2}$$

where we abuse notation on the LHS by using s to denote the random variable which is the realization of the signal. On the RHS, the inner expectation is with respect to uncertainty about θ (conditional on the realization of s) and the outer expectation is with respect to uncertainty about s. One measure of the value of the signal is the difference in these expected payoffs, i.e.,

$$V_{\mathbf{D},p}(\sigma) \equiv \mathbb{E}\left[\sup_{a \in A} \mathbb{E}\left[u(a, \theta) \mid s\right]\right] - \sup_{a \in A} \mathbb{E}\left[u(a, \theta)\right]$$

where $\mathbf{D} = (A, u)$ is the decision problem and $p \in \Delta(\Theta)$ is the agent's prior.

Remark 4.2. It is without loss to assume the use of pure strategies above, but in the subsequent development of the Blackwell order it will be useful to replace a with a mixed strategy $\alpha \in \Delta(A)$ in (4.1) and α with a stochastic map $\alpha : S \to \Delta(A)$ in (4.2).

Example 4.4. Suppose $\Theta = \{\theta_1, \theta_2\}$ with a uniform prior p. The decision problem is (A, u) where $A = \{a_1, a_2\}$ and the utility function $u : A \times \Theta \to \mathbb{R}$ assigns a payoff of 1 when the action matches the state, and zero otherwise. The signal σ is

	s_1	s_2
θ_1	q	$1-q$
θ_2	$1-q$	q

where $q > 1/2$. Then the agent's ex-ante payoff is maximized by choosing the strategy α that maps s_1 to action a_1 and s_2 to action a_2, with an expected payoff of q. In the absence of information the agent's best payoff is $1/2$, so $V_{\mathbf{D},p}(\sigma) = q - 1/2$.

Example 4.5. Suppose $\Theta = \mathbb{R}$ with a prior $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$. The decision problem is (A, u) where $A = \mathbb{R}$ and $u(a, \theta) = -(a - \theta)^2$. The signal is $X = \theta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is independent of θ. Then the agent's ex-ante payoff is maximized by choosing the strategy $\alpha(x) = \mathbb{E}(\theta \mid X = x)$, with an expected payoff of

$$\mathbb{E}_X\left[-(\mathbb{E}(\theta \mid X) - \theta)^2\right] = \mathbb{E}_X\left[-\operatorname{Var}(\theta \mid X)\right] = -\frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}$$

using Fact 2.2 in the final equality (and in particular, the property that posterior variance is independent of the signal realization). In the absence of information the agent's best payoff is $-\operatorname{Var}(\theta) = -\sigma_\theta^2$, so $V_{\mathbf{D},p}(X) = \frac{\sigma_\theta^2\sigma_\varepsilon^2}{\sigma_\theta^2+\sigma_\varepsilon^2} - \sigma_\theta^2 = \frac{\sigma_\theta^4}{\sigma_\theta^2+\sigma_\varepsilon^2}$.

In any specific decision problem, a signal that is informative (in the sense of moving the agent's beliefs about θ) may nevertheless have no instrumental value, as demonstrated in the following exercise.

Exercise 4.4 (U). *Suppose $\Theta = \{1, 2\}$ and let the prior p assign equal probability to either state. Consider the decision problem (A, u) with $A = \{1, 2\}$ and $u(a, \theta) = \mathbb{1}(a = \theta)$. Let σ_P and σ_Q respectively be the two signals described by P and Q in Example 4.1. Show that $V_{\mathbf{D},p}(\sigma_P) = V_{\mathbf{D},p}(\sigma_Q)$. That is, the second independent observation of signal P has no value to the agent over the first.*

4.2.1 Uniformly Better

We'll say that a signal is more informative than another if it is more useful in every decision problem and for every prior belief.

Definition 4.5. *The signal σ is more informative than σ' if $V_{\mathbf{D},p}(\sigma) \geq V_{\mathbf{D},p}(\sigma')$ for every decision problem $\mathbf{D}$ and every prior p.*

This is a strong condition, and we generally won't be able to order signals in this way.

Exercise 4.5 (U). *Let $\Theta = \{\theta_1, \theta_2, \theta_3\}$ with a uniform prior p. Let $A = \{a_1, a_2\}$ and consider two utility functions: Let $u : A \times \Theta \to \mathbb{R}$ take value 1 if $(a, \theta) \in \{(a_1, \theta_1), (a_2, \theta_2), (a_2, \theta_3)\}$, and value 0 otherwise. Let $u' : A \times \theta \to \mathbb{R}$ take value 1 if $(a, \theta) \in \{(a_1, \theta_1), (a_2, \theta_2), (a_1, \theta_3)\}$, and value 0 otherwise. Consider the two information structures*

$$\sigma: \begin{array}{ccc} & s_1 & s_2 \\ \theta_1 & 1 & 0 \\ \theta_2 & 0 & 1 \\ \theta_3 & 0 & 1 \end{array} \qquad \sigma': \begin{array}{ccc} & s_1 & s_2 \\ \theta_1 & 1 & 0 \\ \theta_2 & 0 & 1 \\ \theta_3 & 1 & 0 \end{array}$$

Show that $V_{\mathbf{D},p}(\sigma) > V_{\mathbf{D},p}(\sigma')$ where $\mathbf{D} = (A, u)$, but $V_{\mathbf{D}',p}(\sigma) > V_{\mathbf{D}',p}(\sigma')$ where $\mathbf{D}' = (A, u')$, i.e. the agent prefers the first information given payoffs u and the second given payoffs u'.

The definition of uniformly better varies both the decision problem and also the prior, but the additional flexibility due to arbitrary priors is not substantial:

Exercise 4.6 (G). *Prove that if there is a full-support prior $p_0 \in \Delta(\Theta)$ such that*

$$V_{\mathbf{D},p_0}(\sigma) \geq V_{\mathbf{D},p_0}(\sigma') \quad \text{for every decision problem } \mathbf{D}$$

then σ is more informative than σ'.

4.2.2 Feasible Actions

Our third definition says that a signal is more informative if observing the realization of the signal allows the agent to more effectively tailor his action to the state.

Definition 4.6. *Fix any action set A. A conditional distribution over actions $d : \Theta \to \Delta(A)$ is* feasible *under $\sigma : \Theta \to \Delta(S)$ if there exists a mapping $\alpha : S \to \Delta(A)$ such that*

$$d(a \mid \theta) = \int_{s \in S} \alpha(a \mid s) \sigma(s \mid \theta) ds$$

We'll use $\Lambda_\sigma(A)$ to denote the set of all feasible distributions under σ given action set A.

When σ is a fully revealing signal (e.g., $\sigma : \Theta \to \Delta(\Theta)$ satisfying $\sigma(\theta \mid \theta) = 1$ for every θ), then every mapping $d : \Theta \to \Delta(A)$ is feasible under σ. (Simply set $\alpha = d$.) When σ is uninformative—for example, a constant—then $\Lambda_\theta(A)$ consists of all mappings $d : \Theta \to \Delta(A)$ that take each state into the same distribution over actions. Larger sets $\Lambda_\sigma(A)$ allow the agent more flexibility in tailoring his action to the state, and in this sense are more valuable.

Remark 4.3. Observe that α is itself a Markov kernel, so d can be interpreted as a garbling of σ where A is the set of signal realizations.

4.3 Dispersion of Posterior Beliefs

Our final perspective adopts the view on a signal introduced in Section 2.2, where a signal is identified with the distribution over posterior beliefs that it induces. We consider the dispersion of these posterior beliefs. Given an uninformative signal, the agent's posterior is deterministically equal to the agent's prior, so there is no dispersion. And if the signal reveals the state directly, then the posterior belief is a point mass on the true state, which "maximally varies" depending on the realization of the signal.

We may expect more informative signals to be associated with more dispersed beliefs, but the measure of dispersion is important. For example, using variance to measure dispersion yields a complete order on signals, which cannot possibly be equivalent to the (strict) partial order described in the previous definitions. Below we define two alternative measures for dispersion—mean-preserving spreads and dominance in the convex order—which will turn out to again characterize the previous partial order on signals.

4.3.1 Mean-Preserving Spreads

Definition 4.7. *A distribution of posterior beliefs $F \in \Delta(\Delta(\Theta))$ is a* mean-preserving spread *of another distribution $\widetilde{F}$ if there exist $\Delta(\Theta)$-valued random variables $Z, \widetilde{Z}$ satisfying the following conditions:*

1. $Z \sim F, \widetilde{Z} \sim \widetilde{F}$
2. $\mathbb{E}(Z \mid \widetilde{Z}) = \widetilde{Z}$ *(thus in particular* $\mathbb{E}(Z) = \mathbb{E}(\widetilde{Z})$*)*

The name "mean-preserving spread" reflects that each realization of $\widetilde{Z}$ is spread out into a random Z with the same mean. When Z and $\widetilde{Z}$ are both real-valued, then the second condition can also be stated as $Z = \widetilde{Z} + \varepsilon$ for some random variable ε satisfying $\mathbb{E}(\varepsilon \mid \widetilde{Z}) = 0$.

Example 4.6. Consider the two signals

$$P = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \qquad Q = \begin{pmatrix} 9/16 & 3/16 & 3/16 & 1/16 \\ 1/16 & 3/16 & 3/16 & 9/16 \end{pmatrix}$$

from Example 4.1, where the set of states is $\Theta = \{\theta_1, \theta_2\}$. Let the agent's prior be uniform over these states. Then the agent has two possible posterior beliefs after observing P, $(3/4, 1/4)$ and $(1/4, 3/4)$, which are equally likely. We will write the distribution of posterior beliefs as

$$F_P = 1/2 \cdot (3/4, 1/4) + 1/2 \cdot (1/4, 3/4).$$

Under Q, the distribution of posterior beliefs is instead

$$F_Q = 5/16 \cdot (9/10, 1/10) + 3/8 \cdot (1/2, 1/2) + 5/16 \cdot (1/10, 9/10).$$

We will now show that F_Q is a mean-preserving spread of F_P. Let $\widetilde{Z}$ be a random variable satisfying $\widetilde{Z} \sim F_P$ and construct the random variable Z given $\widetilde{Z}$ as follows:

- If $\widetilde{Z} = (1/4, 3/4)$ then $Z = (1/10, 9/10)$ with probability 5/8 and $Z = (1/2, 1/2)$ with probability 3/8.

- If $\widetilde{Z} = (3/4, 1/4)$ then $Z = (9/10, 1/10)$ with probability 5/8 and $Z = (1/2, 1/2)$ with probability 3/8.

Then $\mathbb{E}(Z \mid \widetilde{Z}) = \widetilde{Z}$ and also $Z \sim F_Q$, so F_Q is a mean-preserving spread of F_P as desired. This construction is depicted in Figure 4.1.

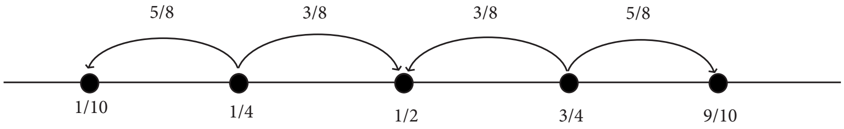

Figure 4.1: Depiction of the mean-preserving spread, where the numbers represent the probability of state θ_1.

Exercise 4.7 (G). *Let $\Theta = \{\theta_1, \theta_2\}$ and consider the two signals*

$$P = \begin{pmatrix} 2/3 & 1/3 \\ 1/4 & 3/4 \end{pmatrix} \qquad Q = \begin{pmatrix} 1/3 & 1/2 & 1/6 \\ 1/8 & 1/2 & 3/8 \end{pmatrix}$$

Define F_P to be the distribution of posterior beliefs induced by P and F_Q to be the distribution of posterior beliefs induced by Q. Prove that F_P is a mean-preserving spread of F_Q.

Exercise 4.8 (G). *Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables, and define $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ to be their sample average. Let $n' < n$ and define $\overline{Y}_{n'} = \frac{1}{n'}\sum_{i=1}^{n'} Y_i$. Prove that the distribution of $\overline{Y}_{n'}$ is a mean preserving spread of the distribution of $\overline{Y}_n$.*

4.3.2 Convex Order

Another partial order of dispersion is the following:

Definition 4.8. *A distribution of posterior beliefs $F \in \Delta(\Delta(\Theta))$* dominates *another distribution G* in the convex order *if for every continuous convex function $h : \Delta(\Theta) \to \mathbb{R}$,*

$$\int_{\Delta(\Theta)} h(p)dF(p) \geq \int_{\Delta(\Theta)} h(p)dG(p)$$

This implies that F and G have the same mean (choosing $h(p) = p$) and that F has the larger variance (choosing $h(p) = \|p\|^2$).

You may recall the concept of *second order stochastic dominance*:

Definition 4.9. *For any lotteries F and G, F* second-order stochastically dominates G *if and only if*

$$\int_{\Delta(\Theta)} u(p)dF(p) \geq \int_{\Delta(\Theta)} u(p)dG(p)$$

for every nondecreasing and concave utility function u.

Dominance in the convex order is stronger than SOSD.

Exercise 4.9 (G). *Prove that if F dominates G in the convex order, then G second order stochastically dominates F.*

The converse is not in general true.

Example 4.7. Let G be a distribution uniform on $[1, 2]$ and let F be a point mass at zero. Then G second order stochastically dominates F but F does not dominate G in the convex order.

Intuitively, second-order stochastic dominance confounds changes in the dispersion of the distribution with shifts in the distribution, while dominance in the convex order isolates the former comparison.

4.4 Blackwell's Theorem and Proof

We now state and prove Blackwell (1951)'s theorem, which demonstrates equivalence of these five definitions. For the proof we will work with finite sets (in particular assuming finite Θ) but several parts of the proof extend more generally.

Theorem 4.1. *The following are equivalent:*

1. *σ' is a garbling of σ.*

2. *σ is more informative than σ'.*

3. *$\Lambda_\sigma(A) \supseteq \Lambda_{\sigma'}(A)$ for every finite action set A.*

4. *For any prior on Θ, if we define F and F' to be the distributions of posterior beliefs induced by σ and σ' (under this prior), then F is a mean-preserving spread of F'.*

5. *For any prior on Θ, if we define F and F' to be the distributions of posterior beliefs induced by σ and σ' (under this prior), then F dominates F' in the convex order.*

Several proofs exist for different parts of this result (see e.g., Blackwell (1951) and Leshno and Spector (1992)). Our proof of the equivalence of (1)-(3) below is based on de Oliveira (2019), which presents a particularly simple and elegant argument.

Proof. Throughout, given stochastic mappings $\alpha : X \to \Delta(Y)$ and $\beta : Y \to \Delta(Z)$, let

$$\beta \circ \alpha(z \mid x) \equiv \sum_{y \in Y} \beta(z \mid y)\alpha(y \mid x) \qquad \forall x \in X, z \in Z.$$

($1 \Rightarrow 3$) 1 implies existence of a mapping $\gamma : S \to \Delta(S')$ such that $\gamma \circ \sigma = \sigma'$, as illustrated below:

$$\begin{tikzcd}
\Theta \arrow[r, "\sigma"] \arrow[d, "\sigma'"'] & S \arrow[dl, "\gamma"] \\
S' &
\end{tikzcd}$$

Consider any action set A and mapping $\alpha' : S' \to \Delta(A)$, where $d = \alpha' \circ \sigma'$ is a feasible distribution under σ'. Define $\alpha = \alpha' \circ \gamma$. Then

$$\alpha \circ \sigma = (\alpha' \circ \gamma) \circ \sigma = \alpha' \circ (\gamma \circ \sigma) = \alpha' \circ \sigma' = d$$

using associativity of the operation $\circ$. So d is feasible also under σ, as depicted in the figure below:

$$\begin{tikzcd}
\Theta \arrow[r, "\sigma"] \arrow[d, "\sigma'"'] & S \arrow[dl, "\gamma"] \arrow[d, "\alpha"] \\
S' \arrow[r, "\alpha'"'] & A
\end{tikzcd}$$

($3 \Rightarrow 1$) Let the action set be S' and define α' to be the identity mapping $id_{S'} : S' \to \Delta(S')$ which satisfies $id_{S'}(s') = \delta_{s'}$ for all $s' \in S'$ (where $\delta_{s'}$ denotes a point mass at s'). By 3, there must exist some $\alpha : S \to \Delta(S')$ such that

$$\alpha \circ \sigma = id_{S'} \circ \sigma'$$

The RHS reduces to σ' since for any $s' \in S'$,

$$id_{S'} \circ \sigma'(s' \mid \theta) = \sum_{s \in S'} id_{S'}(s' \mid s)\sigma'(s \mid \theta) = \sigma'(s' \mid \theta).$$

Thus $\alpha \circ \sigma = \sigma'$. But this implies that σ' is a garbling of σ, as depicted below.

$$\begin{tikzcd}
\Theta \arrow[r, "\sigma"] \arrow[d, "\sigma'"'] & S \arrow[d, dashed, "\alpha"] \\
S' \arrow[r, "id_{S'}"'] & S'
\end{tikzcd}$$

(3 $\Rightarrow$ 2) Clear.

(2 $\Rightarrow$ 3) Suppose 3 fails. Then there is a finite action set A and a vector $\lambda' \in \Lambda_{\sigma'}(A)$ such that $\lambda' \notin \Lambda_\sigma(A)$. The set Λ_σ is a compact and convex subset of $\mathbb{R}^{|\Theta|\times|A|}$ (you will be asked to prove this in Exercise 4.12). Thus by the Separating Hyperplane Theorem, there exists a vector $v \in \mathbb{R}^{|\Theta|\times|A|}$ such that for all $\lambda \in \Lambda_\sigma(A)$,

$$\sum v(a,\theta)\lambda(a,\theta) < \sum v(a,\theta)\lambda'(a,\theta) \tag{4.3}$$

as depicted in Figure 4.2.

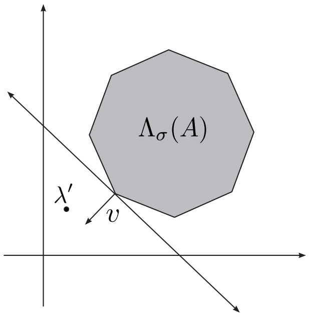

Figure 4.2: Separation of λ' from $\Lambda_\sigma(A)$.

Consider an agent with a uniform prior p on Θ and utility function v, and define $d(a \mid \theta) \equiv \lambda(a,\theta)$ and $d'(a \mid \theta) \equiv \lambda'(a,\theta)$. Then

$$\begin{aligned}\sup_{\alpha:S\to\Delta(A)} \sum_{a,s,\theta} v(a,\theta)\alpha(a \mid s)p(\theta,s) &= \sup_{\alpha:S\to\Delta(A)} \frac{1}{|\Theta|}\sum_{\theta,a,s} \sigma(s \mid \theta)\alpha(a \mid s)v(a,\theta)\\ &= \sup_{d\in\Lambda_\sigma(A)} \frac{1}{|\Theta|}\sum_{\theta,a} d(a \mid \theta)v(a,\theta)\\ &< \frac{1}{|\Theta|}\sum_{\theta,a} d'(a \mid \Theta)v(a,\theta)\end{aligned}$$

using (4.3) in the final inequality. Thus there is a decision problem and a prior for which an agent can achieve a strictly higher payoff by conditioning on σ' rather than on σ, and so 2 fails.

(1 $\Rightarrow$ 4) Let X and X' respectively denote the random realizations of the signals σ and σ'. Since by assumption σ' is a garbling of σ, we can generate θ, X, X' in a way such that X' is independent of θ conditional on X (see Remark 4.1). On this probability space, define Z to be the random posterior belief of θ given X, i.e., the distribution of $\theta \mid X$, and define Z' to be the random posterior belief of θ given X', i.e., the distribution of $\theta \mid X'$. We need to show that $\mathbb{E}[Z \mid Z'] = Z'$.

For any realization θ_i of θ, define $Z_i \equiv \mathbb{E}[\mathbb{1}_{\theta_i} \mid X] = \mathbb{E}[\mathbb{1}_{\theta_i} \mid X, X']$ (where the second equality is due to independence of θ and X' conditional on X) and define $Z_i' \equiv \mathbb{E}[\mathbb{1}_{\theta_i} \mid X']$. Then

$$\begin{aligned}\mathbb{E}[Z_i \mid X'] &= \mathbb{E}[[\mathbb{1}_{\theta_i} \mid X, X'] \mid X'] \\ &= \mathbb{E}[\mathbb{1}_{\theta_i} \mid X'] \\ &= Z_i' \end{aligned} \tag{4.4}$$

where the second equality follows from the law of iterated expectations (henceforth abbreviated to L.I.E.). Moreover,

$$\begin{aligned}\mathbb{E}[Z_i \mid Z'] &= \mathbb{E}[\mathbb{E}[Z_i \mid X', Z'] \mid Z'] && \text{by L.I.E.} \\ &= \mathbb{E}[\mathbb{E}[Z_i \mid X'] \mid Z'] && \text{since } Z' \text{ is a function of } X' \\ &= \mathbb{E}[Z_i' \mid Z'] && \text{using (4.4)} \\ &= Z_i' \end{aligned}$$

Repeating this argument for every θ_i, we have the desired result.

(4 ⇒ 5) Suppose F is a MPS of F' with associated random variables Z and Z' satisfying $\mathbb{E}(Z \mid Z') = Z'$. Then for any continuous and convex function $h : \Delta(\Theta) \to \mathbb{R}$,

$$\begin{aligned}\int_{\Delta(\Theta)} h(p)dF(p) &= \mathbb{E}[h(Z)] \\ &= \mathbb{E}[\mathbb{E}[h(Z) \mid Z']] && \text{by L.I.E.} \\ &\geq \mathbb{E}[h(\mathbb{E}[Z \mid Z'])] && \text{by Jensen's inequality} \\ &= \mathbb{E}[h(Z')] && \text{by assumption of MPS} \\ &= \int_{\Delta(\Theta)} h(p)dF'(p)\end{aligned}$$

So F dominates F' in the convex order.

(5 ⇒ 2) Fix any action set A and utility function u, and define $h : \Delta(\Theta) \to \mathbb{R}$ by

$$h(p) = \max_{a \in A} \sum_{\theta \in \Theta} p(\theta)u(a, \theta)$$

to be the maximum achievable payoff under belief p. The function h is the pointwise maximum of linear functions, and hence it is continuous and convex. Letting $p \sim F$ denote the agent's posterior belief, the maximum *ex-ante* payoff is

$$\int_{\Delta(\Theta)} h(p)dF(p)$$

So dominance in the convex order implies "more valuable." ■

4.5 Additional Exercises

EXERCISE 4.10 (U). *The state space is* $\Theta = \{\theta_1, \theta_2, \theta_3\}$ *and an agent's prior belief is* $p = (1/2, 1/4, 1/4)$. *The agent chooses from actions in the set* $A = \{a_1, a_2, a_3\}$ *and* *has the payoff function*

$$u(a,\theta) = \begin{cases} 1 & \textit{if } a = \theta \\ 0 & \textit{otherwise} \end{cases}$$

(a) *What is the highest expected payoff that the agent can achieve?*

(b) *Now suppose the agent gets to see the outcome of the following signal structure before choosing an action:*

	s_1	s_2
θ_1	0	1
θ_2	$1/2$	$1/2$
θ_3	1	0

(i) *Suppose the realized signal outcome is s_1. Solve for the agent's posterior belief and optimal action.*

(ii) *Suppose the realized signal outcome is s_2. Solve for the agent's posterior belief and optimal action.*

(iii) *What is the agent's best expected payoff (where the expectation is taken prior to the realization of the signal outcome)?*

Exercise 4.11 (U). *(This problem is based on Meyer (1991).) Consider the setting of Example 4.4. It turns out that we can make the second realization of P strictly valuable again by biasing it in favor of the more likely signal realization. That is, let the realizations of P be denoted s_1 and s_2, where*

$$P = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}$$

and modify the second signal in the following way: If the first realization is s_1, then the second signal realization is determined by

$$Q_1 = \begin{pmatrix} 3/4 + c & 1/4 - c \\ 1/4 + c & 3/4 - c \end{pmatrix}$$

and if the first realization is s_2, the second signal realization is determined by

$$Q_2 = \begin{pmatrix} 3/4 - c & 1/4 + c \\ 1/4 - c & 3/4 + c \end{pmatrix}$$

where in both cases the realization of the second signal is independent of the first conditional on the state.

(a) *Show that for any $c \in (0, 1/4]$, the value of observing this second (biased) signal is strictly positive.*

(b) *Solve for the size of the bias $c \in (0, 1/4]$ that leads to the highest expected payoff for the agent.*

Exercise 4.12 (G). *Let the sets A, Θ, and S be finite, and prove that the set $\Lambda_\sigma(A)$ (from Definition 4.6) is compact and convex for every $\sigma : \Theta \to \Delta(S)$.*

Exercise 4.13 (G). *Consider two random variables $X = \theta + \varepsilon$ and $Y = \theta + \varepsilon'$, where θ, ε, and ε' are mutually independent.*

(a) *Suppose that $\theta \sim \mathcal{N}(0,1)$ and $\varepsilon, \varepsilon' \in \mathbb{R}$ are distributed $(\varepsilon, \varepsilon') \sim \mathcal{N}(\mu, \Sigma)$. Prove that X and Y are Blackwell comparable for all mean vectors μ and covariance matrices Σ.*

(b) *Suppose that*

$$\theta \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

and $\varepsilon, \varepsilon' \in \mathbb{R}^2$ are distributed $(\varepsilon, \varepsilon') \sim \mathcal{N}(\mu, \Sigma)$. Prove that X and Y are not always Blackwell ranked by demonstrating a pair (μ, Σ) such that X allows for a strictly higher expected payoff for one decision problem, and Y allows for a strictly higher expected payoff given another.

Exercise 4.14 (G). *In each of the following parts, determine whether the statement is true or false and prove your claim in either case.*

(a) *The state θ belongs to $\{\theta_1, \theta_2\}$ and the two signals are defined as*

$$X = \theta + \varepsilon_1, \quad \varepsilon_1 \sim U([-1/2, 1/2])$$
$$\widetilde{X} = \theta + \varepsilon_2, \quad \varepsilon_2 \sim U([-1/3, 1/3])$$

where U denotes the uniform distribution. The signals X and $\widetilde{X}$ can be Blackwell ranked.

(b) *The state θ belongs to $\{0, 1/3, 2/3, 1\}$ and the two signals are defined as*

$$X = \theta + \varepsilon_1, \quad \varepsilon_1 \sim U([-1/2, 1/2])$$
$$\widetilde{X} = \theta + \varepsilon_2, \quad \varepsilon_2 \sim U([-1/3, 1/3])$$

The signals X and $\widetilde{X}$ can be Blackwell ranked.

Exercise 4.15 (G). *(This problem is based on Brooks, Frankel and Kamenica (2022*a*).) Consider the following strengthening of the Blackwell order. Let θ, X, and X' be random variables defined on the same probability space (Ω, Σ, P).*

Definition 4.10. *Say that X* strongly Blackwell dominates *X' if $(X, \widetilde{X})$ Blackwell dominates $(X', \widetilde{X})$ for every random variable $\widetilde{X}$ also defined on (Ω, Σ, P).*

*Clearly a necessary condition is for X to Blackwell dominate X' (choose $\widetilde{X}$ to be null information). A sufficient condition is for the realization of X' to be known from the realization of X', i.e., for the distribution of $X' \mid X$ to be degenerate for every realization of X (what Brooks, Frankel and Kamenica (2022*b*) call the* refinement order*). Provide an example in between, namely a signal X that strongly Blackwell dominates X', where the realization of X' is not known from X.*

Chapter 5

Comparing Information II: Cost of Information

So far we have considered decision problems in which the signal informing the agent's decision is given exogenously. In many economic applications, agents can acquire information at a cost and thereby control the signal that they observe. The full problem the agent faces is often specified as

$$\max_{\sigma:\Theta\to\Delta(S)} \int_{\Delta(\Theta)} \max_{a\in A} \mathbb{E}_q[u(a,\theta)]d\tau_\sigma(q) - \text{cost of acquiring } \sigma$$

where τ_σ denotes the distribution over posterior beliefs induced by signal σ.

This chapter discusses how to model the cost of information, and is divided into two sections. Section 5.2 considers *prior-dependent* cost functions that are a function both of the agent's prior $p \in \Delta(\Theta)$ and of the signal $\sigma : \Theta \to \Delta(S)$. Section 5.3 considers *prior-independent* cost functions that depend only the signal σ. The former are often interpreted as costs of information processing while the latter are often associated with a physical or exogenous cost of producing information. Both approaches draw from information theory, and we review relevant concepts in Section 5.1.

Two useful benchmarks to keep in mind are the following.

Example 5.1 (Binary). The unknown state θ is equally likely to take the value 0 or 1, and the agent chooses an action $a \in \{0,1\}$ with payoff $u(a,\theta) = \mathbb{1}(a=\theta)$. This action is based on the signal

	$s=0$	$s=1$
$\theta=0$	φ	$1-\varphi$
$\theta=1$	$1-\varphi$	φ

where the agent chooses $\varphi \in [0,1]$.

Example 5.2 (Gaussian). An agent chooses an action $a \in \mathbb{R}$ and receives the payoff $-(a-\theta)^2$, where $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ is an unknown state. This action is based on a signal $X = \theta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and the signal noise σ_ε^2 is chosen by the agent.

5.1 Information Theoretic Preliminaries

This section reviews the definitions of entropy and KL divergence.

5.1.1 Entropy

First assume a finite set of states Θ with $n \equiv |\Theta|$, and consider beliefs $p = (p_1, \ldots, p_n)$ defined over this set.

Definition 5.1 (Shannon (1948)). *Let* $\Theta = \{\theta_1, \ldots, \theta_n\}$ *for any* $n < \infty$. *The* entropy *of belief* $p \in \Delta(\Theta)$ *is*

$$H(p) = -\sum_{\theta \in \Theta} p(\theta) \ln(p(\theta)) = \mathbb{E}_{\theta \sim p}[-\ln(p(\theta))]$$

where $0 \ln 0 = 0$.

Remark 5.1. Entropy is also sometimes defined as a function of the random variable rather than its distribution, i.e., $H(\theta) = \mathbb{E}[-\ln(p(\theta))]$.

Entropy is a quantification of uncertainty in a distribution. The higher the entropy of the distribution, the more information is contained in the realization of a random variable it governs. (Entropy is also often interpreted as the "surprise factor" of the outcome.)

Example 5.3. Suppose $\Theta = \{\theta_1, \theta_2\}$. The entropy of any belief $(q, 1-q)$ is

$$H(q) = -q \ln(q) - (1-q) \ln(1-q). \tag{5.1}$$

This curve is depicted in Figure 5.1 below. It is concave, minimized at the two degenerate distributions $(0,1)$ and $(1,0)$, and maximized at the uniform distribution $(1/2, 1/2)$.

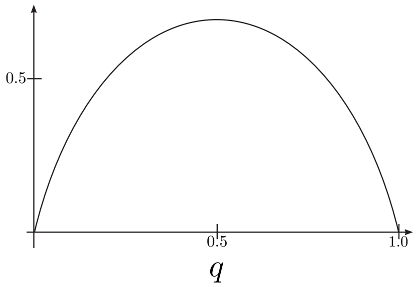

Figure 5.1: Plot of the entropy of the distribution $(q, 1-q)$ as q varies in $[0,1]$.

Several key properties of entropy are collected below.

Property 1 (Maximal Value). $H(p) \leq H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ for every $n < \infty$ and $p \in \Delta(\{\theta_1, \ldots, \theta_n\})$; that is, entropy is maximized at the uniform distribution.

Property 2 (Probability Zero States). $H(p) = H(p_1, \dots, p_n, 0)$ for every $n < \infty$ and $p \in \Delta(\{\theta_1, \dots, \theta_n\})$; that is, entropy is unchanged by an expansion of the state space to include probability-zero outcomes.

Property 3 (Continuity). H is continuous with respect to all of its arguments.

Property 4 (Chain Rule). Suppose $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_m\}$, where the joint distribution of (X, Y) is denoted p, the marginal distribution of X is p_X, and the conditional distribution of Y given X is $p_{Y|X}$. Then

$$H(p) = H(p_X) + \sum_{i=1}^{n} p_X(x_i) H(p_{Y|X=x_i})$$

or more simply

$$H(X, Y) = H(X) + H(Y \mid X)$$

where $H(X, Y) \equiv H(p)$ is the entropy of the joint distribution, $H(X) \equiv H(p_X)$ is the entropy of of the marginal distribution of X, and

$$H(Y \mid X) \equiv \sum_{i=1}^{n} p_X(x_i) H(p_{Y|X=x_i})$$

is the expected entropy of the conditional distribution of Y given X, also known as the *conditional entropy* of Y given X.

REMARK 5.2. In the special case where X and Y are independent, Property 4 implies $H(X, Y) = H(X) + H(Y)$.

Property 5 (Nonnegativity). $H(p) \geq 0$ for all distributions p.

Property 6 (Degenerate Distributions). $H(p) = 0$ for all degenerate distributions p.

Property 7 (Concavity). H is concave.

Property 8 (Relabelling of States). $H(p_1, \dots, p_n) = H(p_{\pi(1)}, \dots, p_{\pi(n)})$ for any bijection π from $\{1, \dots, n\}$ to itself; that is, entropy is invariant to a relabelling of states.

Property 9 (Information Reduces Uncertainty). $H(Y \mid X) \leq H(Y)$ with equality if and only if X and Y are independent; that is, conditioning on information reduces expected entropy.

Properties 1-4 constitute a set of necessary and sufficient conditions for the form of H given in (5.1), up to rescaling.

Proposition 14 (Khinchin (1957)). *Let $H(p_1,\dots,p_n)$ be a function defined for any $n \in \mathbb{Z}_+$ and for all values $p_1,\dots,p_n$ satisfying $p_i \geq 0$ for each $i = 1,\dots,n$ and $\sum_{i=1}^n p_i = 1$. Then H satisfies Properties 1-4 if and only if*

$$H(p_1,\dots,p_n) = -\lambda \sum_{i=1}^{n} p_i \ln(p_i)$$

for some constant $\lambda > 0$.[1]

Properties 5, 6, and 8 are immediate from the functional form of entropy. Property 7 (concavity) follows because $-x\log(x)$ is concave, and the sum of concave functions is concave. (In fact, the same argument shows that entropy is *strictly* concave, so Property 1 can be strengthened to the statement that the uniform distribution is the unique maximum.) The following exercise asks you to prove that entropy satisfies Property 9.

Exercise 5.1 (G). *Suppose $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$, where p_X and p_Y denote the marginal distributions of X and Y, and $p_{Y|X}$ denotes the conditional distribution of Y given X. Let $H(Y) \equiv H(p_Y)$ be the entropy of of the marginal distribution of Y, and $H(Y \mid X) \equiv \sum_{i=1}^n p_X(x_i) H(p_{Y|X=x_i})$ be the conditional entropy of Y given X. Prove that $H(Y \mid X) \leq H(Y)$.*

Shannon (1948) defines a continuous version of entropy.

Definition 5.2. *The entropy of probability density p on $\Theta \subseteq \mathbb{R}$ is*

$$H(p) = -\int_{\theta\in\Theta} p(\theta)\ln(p(\theta))d\theta$$

Example 5.4. Recall that the normally distributed variable $\theta \sim \mathcal{N}(\mu, \sigma^2)$ has density $p(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}$. The entropy of this distribution is

$$\begin{aligned}\mathbb{E}\left[-\ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}\right)\right] &= -\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{1}{2\sigma^2}\mathbb{E}\left[(\theta-\mu)^2\right] \\ &= \frac{1}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2} \qquad (5.2)\end{aligned}$$

using in the second equality that $\mathbb{E}[(\theta-\mu)^2] = \sigma^2$. So entropy and variance order normal distributions in the same way.

5.1.2 Kullback-Liebler Divergence

The *Kullback-Liebler Divergence (KL divergence)*, also known as *relative entropy*, quantifies how different two distributions are.

[1] Recalling that $\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$ for any two bases $a, b > 0$, changing the logarithm to a different basis simply rescales the measure. Choice of base 2 and of base e are both common.

Definition 5.3 (KL-Divergence). *Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ for any $n < \infty$, and let $p, q \in \Delta(\Theta)$. Then the KL divergence from q to p is*

$$D(p\|q) = \sum_{\theta \in \Theta} p(\theta) \ln\left(\frac{p(\theta)}{q(\theta)}\right) = \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{p(\theta)}{q(\theta)}\right)\right]$$

where $0 \ln 0 = 0$.

Example 5.5 (Binary). Let $\Theta = \{\theta_1, \theta_2\}$ with $(p, 1-p)$ and $(q, 1-q)$ be two distributions on this set. Then

$$D(p\|q) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

Intuitively, larger log likelihood ratios $\ln\left(\frac{p}{q}\right)$ and $\ln\left(\frac{1-p}{1-q}\right)$ reflect distributions that are more different. KL divergence aggregates these log likelihood ratios by weighting them with respect to their probabilities under a reference distribution, which is chosen to be either of p or q.

Example 5.6 (Gaussian). Let p and q denote two Gaussian densities with common variance σ and different means μ_p and μ_q. Then

$$\begin{aligned} D(p\|q) &= \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{e^{-\frac{1}{2}\left(\frac{\theta - \mu_p}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{\theta - \mu_q}{\sigma}\right)^2}}\right)\right] \\ &= \frac{\mu_q^2 - \mu_p^2}{2\sigma^2} - \frac{\mu_q - \mu_p}{\sigma^2} \cdot \mathbb{E}_{\theta \sim p}(\theta) = \frac{(\mu_q - \mu_p)^2}{2\sigma^2} \end{aligned}$$

So as we might expect, the further the two means, the larger the KL divergence between the two distributions.

KL divergence is not in general symmetric (with Example 5.6 being a notable exception) and hence it is not a metric. Other key properties of the KL divergence include:

Property 10 (Nonnegativity). $D(p\|q) \geq 0$ for all $p, q \in \Delta(\Theta)$, with equality if and only if $p = q$.

To prove this, observe that

$$\begin{aligned} -D(p\|q) &= \mathbb{E}_{\theta \sim p}\left[\ln\left(\frac{q(\theta)}{p(\theta)}\right)\right] \\ &\leq \ln\left(\mathbb{E}_{\theta \sim p}\left[\frac{q(\theta)}{p(\theta)}\right]\right) && \text{by Jensen's inequality} \\ &= \ln(1) = 0 && \text{since } \textstyle\sum_{\theta \in \Theta} p(\theta)\left(\frac{q(\theta)}{p(\theta)}\right) = 1 \end{aligned}$$

Property 11 (Additivity for Independent Distributions). Suppose $p_1 \in \Delta(\mathcal{X}_1)$ and $p_2 \in \Delta(\mathcal{X}_2)$ are independent distributions, with $p(x_1, x_2) = p_1(x_1)p_2(x_2)$. Likewise suppose $q_1 \in \Delta(\mathcal{X}_1)$ and $q_2 \in \Delta(\mathcal{X}_2)$ are independent distributions with $q(x_1, x_2) = q(x_1)q(x_2)$. Then

$$D(p\|q) = D(p_1\|q_1) + D(p_2\|q_2),$$

i.e., KL divergence is additive for independent distributions.

This property follows from straightforward algebra:

$$\begin{aligned}
D(p\|q) &= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} p(x_1,x_2)\ln\left(\frac{p(x_1,x_2)}{q(x_1,x_2)}\right)\\
&= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} p_1(x_1)p_2(x_2)\ln\left(\frac{p_1(x_1)p_2(x_2)}{q_1(x_1)q_2(x_2)}\right)\\
&= \sum_{x_2\in\mathcal{X}_2} p_2(x_2)\left(\sum_{x_1\in\mathcal{X}_1} p_1(x_1)\ln\left(\frac{p_1(x_1)}{q_1(x_1)}\right)\right)\\
&\quad + \sum_{x_1\in\mathcal{X}_1} p_1(x_1)\left(\sum_{x_2\in\mathcal{X}_2} p_2(x_2)\ln\left(\frac{p_2(x_2)}{q_2(x_2)}\right)\right) = D(p_1\|q_1)+D(p_2\|q_2)
\end{aligned}$$

where independence is invoked in the second equality.

Property 12 (Convexity). D is convex: For any two pairs (p,q) and (p',q'), and any $\alpha\in[0,1]$, we have

$$D\left(\alpha p+(1-\alpha)p'\|\alpha q+(1-\alpha)q'\right)\leq \alpha D(p\|q)+(1-\alpha)D(p'\|q')$$

Exercise 5.2 (G). *Prove the above property using the following fact:*

Fact 5.1 (Log-Sum Inequality). *Let $a_1,\dots a_n$ and $b_1,\dots b_n$ be nonnegative real numbers. Then*

$$\sum_{i=1}^n a_i\ln\left(\frac{a_i}{b_i}\right)\geq\left(\sum_{i=1}^n a_i\right)\ln\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right).$$

There is a close relationship between KL divergence and entropy. First, the entropy of a distribution $p\in\Delta(\Theta)$ with $n\equiv|\Theta|<\infty$ can be rewritten directly in terms of KL divergence:

$$H(p)=\ln n-D(p\|U)$$

where U denotes the uniform distribution on Θ. Thus, the larger the KL divergence from the uniform distribution to p, the lower the entropy of p. This is proved by observing that

$$\begin{aligned}
\ln n-D(p\|U) &= \ln n-\sum_{\theta\in\Theta}p(\theta)\ln\left(\frac{p(\theta)}{1/n}\right)\\
&= \sum_{\theta\in\Theta}p(\theta)(\ln n-\ln(np(\theta))) \qquad \text{since } \textstyle\sum_{\theta\in\Theta}p(\theta)=1\\
&= -\sum_{\theta\in\Theta}p(\theta)\ln(p(\theta))=H(p)
\end{aligned}$$

Remark 5.3. Together with Property 10, the above relationship implies that entropy is maximized at the uniform distribution (Property 1).

KL divergence cannot be rewritten directly in terms of entropy, although

$$D(p\|q)=-\sum_{\theta\in\Theta}p(\theta)\ln(q(\theta))-H(p)$$

where $-\sum_{\theta\in\Theta}p(\theta)\ln(q(\theta))$ is the *cross-entropy* of distribution q relative to p.

5.2 Prior-Dependent Costs

Returning to the question of how to model the cost function, we begin with *prior-dependent* cost functions. Dependence on the prior belief means that the cost of absorbing the information content of a signal varies with what the agent already knows. This feature may be justified if we view the cost of information as an information processing or cognitive cost: For example, processing a news article about a proposed tax change may be relatively easy for someone who already understands this tax change well, but cognitively taxing for someone who does not.

It will be convenient to represent signals as distributions over posterior beliefs, as in Section 2.2.2. Following Definition 2.2, we use $\mathcal{T}(p)$ to denote the set of Bayes plausible distributions given prior p, and we further define

$$\mathcal{S} = \{(p, \tau) : p \in \Delta(\Theta), \tau \in \mathcal{T}(p)\}$$

to be the domain of prior beliefs and Bayes plausible distributions. The cost functions in this section will take the form $C : \mathcal{S} \to \mathbb{R}$.

5.2.1 Uniform Posterior Separability

One popular class of cost functions are those that are *uniformly posterior separable.*

Definition 5.4 (Caplin and Dean (2013); Caplin, Dean and Leahy (2022)). *The cost function* $C : \mathcal{S} \to \mathbb{R}$ *is* uniformly posterior separable *(henceforth UPS) if there is a strictly concave function* Φ *such that*

$$C(p, \tau) = \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)] \quad \forall (p, \tau) \in \mathcal{S}.$$

We can interpret this cost of information as the expected reduction of uncertainty, where $\Phi : \Delta(\Theta) \to \mathbb{R}$ measures how uncertain the belief is.

Remark 5.4. The cost of "no information" is zero, since $\Phi(p) - \mathbb{E}_{q\sim\delta_p}[\Phi(q)] = \Phi(p) - \Phi(p) = 0$ (with δ_p denoting the degenerate distribution at the prior p).

Remark 5.5. Concavity of Φ guarantees that uncertainty decreases in expectation when more information is received. Together with Bayes plausibility of τ, this further implies that UPS cost functions are everywhere positive:

$$\begin{aligned}
\Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)] &\geq \Phi(p) - \Phi(\mathbb{E}_{q\sim\tau}[q]) && \text{by Jensen's inequality} \\
&= \Phi(p) - \Phi(p) && \text{by Bayes plausibility of } \tau \\
&= 0
\end{aligned}$$

Remark 5.6. UPS cost functions are consistent with the Blackwell order. That is, let σ and σ' be arbitrary signals where σ Blackwell dominates σ'. Fix any prior p, and let τ_σ and $\tau_{\sigma'}$ denote the distributions over posteriors that are induced by σ and σ'. Then for any UPS cost function C, we have $C(p, \tau_\sigma) \geq C(p, \tau_{\sigma'})$ since

$$C(p, \tau) = \int (\Phi(p) - \Phi(q)) d\tau(q)$$

where $\Phi(p) - \Phi(q)$ is convex in q, and τ_σ dominates $\tau_{\sigma'}$ in the convex order (see the characterization of the Blackwell order in Section 4.3.2).

The leading specification of C is the expected reduction of the entropy of the agent's belief.

Example 5.7 (Entropy Reduction). Let H be the entropy function given in Definition 5.1. Then define

$$C_{\text{Ent}}(p,\tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)] \quad \forall (p,\tau) \in \mathcal{S} \tag{5.3}$$

to be the expected reduction in the entropy of the agent's belief.

Initially proposed as an information cost in Sims (2003), this cost function is a cornerstone of the rational inattention literature (Caplin and Dean, 2013; Caplin, Dean and Leahy, 2015; Hebert and Woodford, 2021*a*; Hébert and La'O, 2022). Various conceptual foundations for entropic costs and uniformly posterior separable cost functions (as well as the broader class of posterior separable cost functions discussed in Section 5.2.3) can be found in Caplin and Dean (2013), Matějka and McKay (2015), Morris and Strack (2019), Hebert and Woodford (2021*b*), Bloedel and Zhong (2021), and Denti (2022) among others.

Example 5.8. In the setting of Example 5.1, we have

$$C_{\text{Ent}}(p,\tau_\varphi) = -\ln\left(\frac{1}{2}\right) + (\varphi\ln(\varphi) + (1-\varphi)\ln(1-\varphi))$$

where τ_φ denotes the distribution over posterior beliefs induced by the signal indexed to φ. The cost of the signal is largest when $\varphi \in \{0,1\}$ (corresponding to a fully revealing signal) and smallest when $\varphi = 1/2$ (corresponding to an uninformative signal).

Besides entropy, another natural choice of Φ is variance.

Example 5.9 (Variance Reduction). Let

$$C_{\text{Var}}(p,\tau) = \text{Var}(p) - \mathbb{E}_{q\sim\tau}[\text{Var}(q)] \tag{5.4}$$

be the expected reduction in the variance of the agent's belief.

Exercise 5.3 (G). *Prove that variance is strictly concave, so C_{Var} is a UPS cost function.*

Example 5.10. Consider the setting of Example 5.2 (where we use $\tau_{\sigma_\varepsilon^2}$ to denote the distribution over posterior beliefs induced by observing the signal $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$). Applying (5.2),

$$\begin{aligned} C_{Ent}(p,\tau_{\sigma_\varepsilon^2}) &= \left(\frac{1}{2}\ln(2\pi\sigma_\theta^2) + \frac{1}{2}\right) - \left(\frac{1}{2}\ln\left(2\pi\left(\frac{\sigma_\theta^2\sigma_\varepsilon^2}{\sigma_\theta^2+\sigma_\varepsilon^2}\right)\right) + \frac{1}{2}\right) \\ &= \frac{1}{2}\ln\left(\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_\varepsilon^2}\right) \end{aligned}$$

while

$$C_{Var}(p, \tau_{\sigma_\varepsilon^2}) = \sigma_\theta^2 - \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} = \frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

For every fixed prior variance σ_θ^2, both cost functions are strictly decreasing in the noise variance σ_ε^2, and thus correspond to different cardinal representations of the same ordering over signals. One interesting contrast is that $C_{Ent}(p, \sigma_\varepsilon^2) \to \infty$ as $\sigma_\varepsilon^2 \to 0$, while $C_{Var}(p, \sigma_\varepsilon^2) \to \sigma^2$. That is, the cost of information using C_{Var} is bounded above by the agent's prior uncertainty, while entropy cost is unbounded.

5.2.2 Decision-Theoretic Foundations

The function Φ is interpreted in the previous section as a "pure" measure of uncertainty, without reference to why this uncertainty matters. Parallel to Section 4.2's assessment of the value of information using decision problems, Frankel and Kamenica (2019) microfound the function Φ as measuring the instrumental loss of uncertainty for a specific decision problem.

DEFINITION 5.5. *For any belief $q \in \Delta(\Theta)$ and decision problem $\mathcal{D} = (A, u)$, let*

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q \left[\max_{a \in A} u(a, \theta) \right] - \max_{a \in A} \mathbb{E}_q \left[u(a, \theta) \right].$$

The first term of this expression is the agent's best expected payoff when conditioning his action directly on the realized state (which is random and distributed according to the agent's belief q). The second term is the best expected payoff that the agent with belief q can achieve given no additional information on which to condition his action. Thus $\Phi_{\mathcal{D}}$ quantifies the agent's payoff loss from not knowing a state which is distributed according to q.

DEFINITION 5.6 (Frankel and Kamenica (2019)). *Say that $\Phi : \Delta(\Theta) \to \mathbb{R}$ is* valid *if there is a decision problem $\mathcal{D}$ such that $\Phi = \Phi_{\mathcal{D}}$.*

Any function Φ that is concave and takes value zero at degenerate distributions (i.e., satisfies Properties 6 and 7) can be microfounded using a decision problem in this way.

Proposition 15 (Frankel and Kamenica (2019))**.** *$\Phi : \Delta(\Theta) \to \mathbb{R}$ is valid if and only if it satisfies Properties 6 and 7.*

This result follows from the subsequent lemma, which is of independent interest.

Lemma 1. *Let Θ be a finite set. Then every convex function $V : \Delta(\Theta) \to \mathbb{R}$ can be represented as*

$$V(q) = \sup_{a \in A} \mathbb{E}_q[u(a, \theta)] \quad \forall q \in \Delta(\Theta) \tag{5.5}$$

for some decision problem (A, u), where A is a set (not necessarily finite) and u is a map $u : \Theta \times A \to [-\infty, +\infty]$.

The key points in the proof of this lemma are that $\mathbb{E}_q(u(a,\theta))$ is affine in q, and that every convex function is the supremum of affine functions lying below it. We'll prove this lemma assuming that V is continuous and has a nonvertical supporting hyperplane at every point $q \in \Delta(\Theta)$, leaving the completion of the proof when these assumptions fail as Exercise 5.4.[2]

Proof. Our approach is to construct a set of actions indexed to beliefs, $A = \{a_q : q \in \Delta^n\}$, and to construct a utility function such that each action a_q is optimal at the belief q. To do this, define a family of affine functions $(U_{a_q})_{q\in\Delta^n}$, where each $U_{a_q} : \Delta^n \to \mathbb{R}$ is a supporting hyperplane of the epigraph of V at q, as depicted below in Figure 5.2.[3]

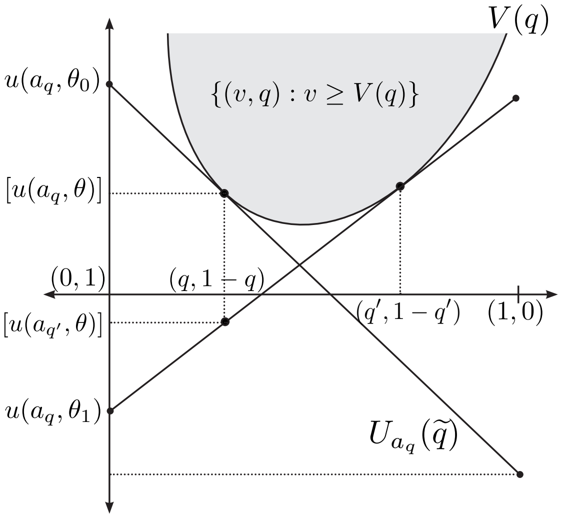
$V(q)$
$u(a_q,\theta_0)$
$\{(v,q) : v \geq V(q)\}$
$\mathbb{E}_q\left[u(a_q,\theta)\right]$
$(0,1)$
$(q,1-q)$
$(q',1-q')$
$(1,0)$
$\mathbb{E}_q\left[u(a_{q'},\theta)\right]$
$u(a_q,\theta_1)$
$U_{a_q}(\widetilde{q})$

Figure 5.2: Example construction for a binary state space $\Theta = \{\theta_0, \theta_1\}$. The action a_q is optimal at belief q; that is, for every other belief q' we have $\mathbb{E}_q[u(a_q,\theta)] \geq \mathbb{E}_q[u(a_{q'},\theta)]$, as depicted here.

Since V is continuous and convex, it can be represented on its domain as the supremum of all affine functions lying below it. Since each U_{a_q} is affine and lies below V, we have that

$$V(q) \geq U_a(q) \quad \forall a \in A, q \in \Delta^n.$$

Moreover (by definition) U_{a_q} supports V at q, so $U_{a_q}(q) = V(q)$. This implies that

$$U_{a_q}(q) = \max_{a\in A} U_a(q) \quad \forall q \in \Delta(\Theta) \tag{5.6}$$

We now need to express U_{a_q} as an expected utility function. Since each belief q' is a convex combination of the degenerate beliefs $(\delta_\theta)_{\theta\in\Theta}$ (with weights given

[2] Under these assumptions, the supremum in (5.5) can be replaced with maximum, as the following proof demonstrates.

[3] Recall that the epigraph of V is $\{(q,v) : v \geq V(q)\}$, the set of points lying on or above V.

by $q'(\theta)$), and U_{a_q} is affine, it follows that

$$U_{a_q}(q') = \sum_{\theta \in \Theta} q'(\theta) U_{a_q}(\delta_\theta) \quad \forall q' \in \Delta(\Theta) \tag{5.7}$$

Now define the utility function $u : \mathbb{R}^n \to \mathbb{R}$ to satisfy $u(a, \theta) = U_a(\delta_\theta)$ for every $a \in A$ and $\theta \in \Theta$. Then from (5.7),

$$U_{a_q}(q') = \sum_{\theta \in \Theta} q'(\theta) u(a_q, \theta)$$

and so (5.6) implies that

$$\mathbb{E}_q[u(a_q, \theta)] \geq \mathbb{E}_q[u(a, \theta)]$$

for every $q \in \Delta(\Theta)$ and $a \in A$. Thus each action a_q is optimal at belief q, and achieves the expected utility $U_{a_q}(q) = V(q)$ as desired. ■

Exercise 5.4 (G). *Complete the proof by showing that the statement of Lemma 1 continues to hold when V is discontinuous and/or there exists a belief q at which every supporting hyperplane of V is vertical.*

Hint 1. *Observe that vertical supporting hyperplanes can only exist on the boundary of $\Delta(\Theta)$, and that discontinuities can only occur at degenerate beliefs.*

We'll now use this lemma to prove Proposition 15.

Proof. Suppose Φ satisfies Assumptions 6 and 7. Then $-\Phi$ is convex, so by Lemma 1, there is a set of actions A and a utility function $u : A \times \Theta \to \mathbb{R}$ such that

$$-\Phi(q) = \max_{a \in A} \mathbb{E}_q[u(a, \theta)] \quad \forall q \in \Delta(\Theta). \tag{5.8}$$

We need to verify that

$$\Phi(q) = \mathbb{E}_q\left[\max_{a \in A} u(a, \theta)\right] - \max_{a \in A} \mathbb{E}_q[u(a, \theta)] \tag{5.9}$$

for every $q \in \Delta(\Theta)$. Again index the states by $\theta_1, \ldots, \theta_n$ (where $n \equiv |\Theta|$), and define δ_{θ_i} to be the belief that is degenerate at state θ_i. Then for any $\theta_i \in \Theta$

$$\begin{aligned}
\max_{a \in A} u(a, \theta_i) &= \max_{a \in A} \mathbb{E}_{\delta_{\theta_i}}[u(a, \theta)] \\
&= -\Phi(\delta_{\theta_i}) && \text{by (5.8)} \\
&= 0 && \text{by Assumption 6}
\end{aligned}$$

Thus also $\mathbb{E}_q\left[\max_{a \in A} u(a, \theta)\right] = 0$ for any belief q, which together with (5.8) implies that (5.9) reduces to $\Phi(a) = 0 - (-\Phi(a))$ and is thus true.

In the other direction,

$$\Phi(\delta_\theta) = \max_{a \in A} u(a, \theta) - \max_{a \in A} u(a, \theta) = 0 \quad \forall \theta \in \Theta$$

implying Property 6. Concavity of Φ (Property 7) follows by construction of Φ since $\mathbb{E}_q\left[\max_{a\in A} u(a,\theta)\right]$ is affine while $\sup_{a\in A} \mathbb{E}_q[u(a,\theta)]$ is a pointwise supremum of affine functions, and thus convex. ■

By Proposition 15, the two example cost functions from the previous section, C_{Ent} and C_{Var}, can be microfounded using decision problems. These decision problems are given below.

Example 5.11 (Microfoundation for Entropy Cost). Set $A = \Delta(\Theta)$ and $u(a,\theta) = \ln(a(\theta))$, where $\ln 0 = -\infty$. Then the cost of uncertainty is

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q\left[\max_a [\ln(a(\theta))]\right] - \max_a \mathbb{E}_q\left[\ln(a(\theta))\right] = H(q).$$

Example 5.12 (Microfoundation for Variance Cost). Set $A = \Theta \subseteq \mathbb{R}$ and $u(a,\theta) = -(a-\theta)^2$. Then

$$\Phi_{\mathcal{D}}(q) = \mathbb{E}_q\left[\max_a \left[-(a-\theta)^2\right]\right] - \max_a \mathbb{E}_q\left[-(a-\theta)^2\right] = Var_q(\theta)$$

5.2.3 Posterior Separability

A weaker requirement than uniform posterior separability is that the cost of τ can be written in a way that is separable in the realized posteriors.

Definition 5.7 (Caplin and Dean (2013); Caplin, Dean and Leahy (2022)). *The cost function $C : S \to \mathbb{R}$ is* posterior separable *if*

$$C(p,\tau) = \mathbb{E}[\Phi_p(q)]$$

for some family of convex functions $(\Phi_p)_{p\in\Delta(\Theta)}$ where each $\Phi_p : \Delta(\Theta) \to \mathbb{R}$ is everywhere weakly positive, and $\Phi_p(p) = 0$ for every p.

Remark 5.7. When the cost function is posterior separable but not uniformly posterior separable, the cost of acquiring two signals in sequence may depend on the order in which these signals are acquired. This is not true for for UPS cost functions (Frankel and Kamenica, 2019; Bloedel and Zhong, 2021).

When the cost function is posterior separable, then the agent's payoff from choosing signal $\sigma : \Theta \to \Delta(S)$ and strategy $\alpha : S \to \Delta(A)$ is

$$\int_{\Delta(\Theta)} \int_{a\in A} \alpha(a \mid q)\mathbb{E}_q[u(a,\theta)]d\tau_\sigma(q) - C(p,\tau_\sigma),$$

and can be rewritten as

$$\int_{\Delta(\Theta)} \int_{a\in A} \alpha(a \mid q)\left(\mathbb{E}_q[u(a,\theta)] - \Phi_p(q)\right) d\tau_\sigma(q)$$

where the concave function $\mathbb{E}_q[u(a,\theta)] - \Phi_p(q)$ is the "net utility" of action a under posterior q. So maximizing the value function is equivalent to maximizing the expected net utility over all Bayes-plausible distributions and strategies, which is an optimization problem that can be solved using standard methods. This tractability is a part of the appeal of this family of cost functions.

A closely related concept appears in Frankel and Kamenica (2019), where $\Phi_p(q)$ is interpreted as the amount of information in news that moves an agent's belief from p to q. Frankel and Kamenica (2019) define the pair (Φ_p, Φ) as *coupled* if $\mathbb{E}[\Phi_p(q)] = \mathbb{E}[\Phi(p) - \Phi(q)]$, in which case the cost function is not only posterior separable but also uniformly posterior separable.

That uniform posterior separability is strictly stronger than posterior separability is nearly immediate, except for the requirement in the definition of posterior separable cost functions that $\Phi_p(q)$ is everywhere positive. We cannot therefore simply convert a UPS cost function $C(p, \tau) = \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)]$ into a posterior separable cost function $C(p, \tau) = \mathbb{E}[\Phi_p(q)]$ by setting $\Phi_p(q) \equiv \Phi(p) - \Phi(q)$, as this quantity may be negative for some posterior beliefs q. The correct construction is instead to choose Φ_p to be a *Bregman divergence* of Φ (Frankel and Kamenica, 2019; Caplin, Dean and Leahy, 2022).

Definition 5.8. *Let* $\Phi : \Delta(\Theta) \to \mathbb{R}$ *be a concave function. A* supergradient *of* Φ *at* $p \in \Delta(\Theta)$ *is any vector* $\nabla\Phi(p)$ *such that*

$$\Phi(p) + \nabla\Phi(p) \cdot (q - p) \geq \Phi(q)$$

for every $q \in \Delta(\Theta)$.

Remark 5.8. When Φ is concave, then a supergradient $\nabla\Phi(q)$ exists for every q. When Φ is smooth at q, then $\nabla\Phi(q)$ is unique and equal to $\Phi'(q)$.

Definition 5.9. *Let* $\Phi : \Delta(\Theta) \to \mathbb{R}$ *be a concave function. A* Bregman divergence *of* Φ *is any map* $D_\Phi : \Delta(\Theta) \times \Delta(\Theta) \to \mathbb{R}$ *satisfying*

$$D_\Phi(p, q) = \Phi(p) - \Phi(q) + \nabla\Phi(p) \cdot (q - p) \quad \forall (p, q) \in \Delta(\Theta) \times \Delta(\Theta)$$

where $\nabla\Phi(q)$ *is a supergradient of* Φ *at* q.

This is the difference between the value of Φ at q and the value of the first-order Taylor expansion of Φ around p evaluated at point q.

Setting $\Phi_p(q) = D_\Phi(p, q)$ from Definition 5.9, we have

$$\Phi_p(q) = (\Phi(p) + \nabla\Phi(p) \cdot (q - p)) - \Phi(q) \geq 0 \quad \forall q \in \Delta(\Theta)$$

since $\nabla\Phi(p)$ is a supergradient of Φ, and also

$$\begin{aligned} \mathbb{E}_{q\sim\tau}[\Phi_p(q)] &= \mathbb{E}_{q\sim\tau}[\Phi(p) - \Phi(q) + \nabla\Phi(p) \cdot (q - p)] \\ &= \Phi(p) - \mathbb{E}_{q\sim\tau}[\Phi(q)] \end{aligned}$$

using in the second inequality that $\mathbb{E}_{q\sim\tau}(q - p) = 0$. The relationship between Φ_p and Φ is depicted in Figure 5.3.

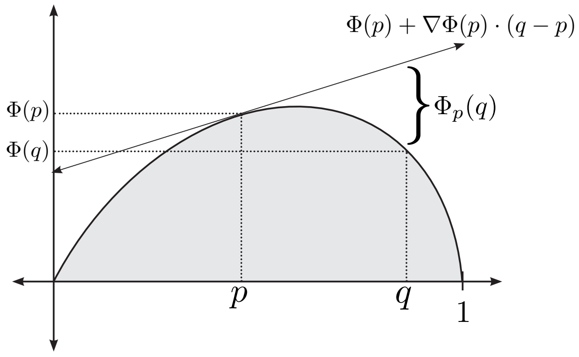

Figure 5.3: Relationship between Φ_p and Φ.

Example 5.13. Consider entropy cost $C_{\text{Ent}}(p, \tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)]$. The Bregman divergence of entropy is KL divergence (Bregman, 1967), so

$$C_{\text{Ent}}(p, \tau) = H(p) - \mathbb{E}_{q\sim\tau}[H(q)] = \mathbb{E}[D(p\|q)].$$

Thus we can view the cost of a signal that generates the distribution of beliefs τ either as the expected reduction in the entropy of the agent's belief, or as the expected KL divergence from the agent's prior to the realized posterior belief.

5.3 Prior-Independent Costs

We now turn to cost functions that do not depend on the agent's prior belief. If the cost of information is exogenous to the agent—for example, a price determined within a market, or a physical cost of producing information—then we may expect the cost of acquiring information to be the same for all consumers regardless of their beliefs or expertise in the area, and thus prior independent.

One common cost specification is the following.

Example 5.14. In the setting of Example 5.2, let

$$C(\sigma_\varepsilon^2) = \frac{\kappa}{\sigma_\varepsilon^2} \tag{5.10}$$

Then the cost of the signal scales linearly with the precision of the signal, $1/\sigma_\varepsilon^2$. This formulation of the cost is especially sensible if we interpret θ as an unknown population parameter (for instance, the average height in a population) and the signal as a sample of individuals from this population. Modeling each observation as $X_i = \theta + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent of θ and independent across agents, the conditional distribution of θ given the sample $(X_1, \dots, X_n)$ is the same as the conditional distribution of θ given the signal $X = \theta + \delta$, $\delta \sim \mathcal{N}(0, \sigma^2/n)$ (see Exercise 2.10). So (5.10) corresponds to a fixed cost of κ/σ^2 for each individual in the sample. This cost function is used in Wald's classic model of sequential sampling (Wald, 1945; Arrow, Blackwell and Girshick, 1949), and is a common modeling choice in continuous-time sequential sampling problems where the signal corresponds to observation of a Brownian motion (Fudenberg, Strack and Strzalecki, 2018; Liang, Mu and Syrgkanis, 2022).

We now present a generalization of the above cost function due to Pomatto, Strack and Tamuz (2020). Let Θ be a finite set and S be a set of signal realizations equipped with σ-algebra Σ, with $\Delta(S)$ denoting the set of measurable probability distributions on S. A signal is a mapping $\sigma : \Theta \to \Delta(S)$, and we use $\sigma_\theta \equiv \sigma(\cdot \mid \theta) \in \Delta(S)$ to denote the conditional distribution over signal realizations when the state is θ.

Definition 5.10. *The log-likelihood ratio between states θ and θ' at signal realization s is*

$$\ell^{\sigma}_{\theta,\theta'}(s) = \ln\left(\frac{d\sigma_\theta(s)}{d\sigma_{\theta'}(s)}\right)$$

Definition 5.11. *For any state $\theta \in \Theta$ and map $\alpha : \Theta \to \mathbb{N}$, define*

$$M^{\sigma}_{\theta}(\alpha) = \int_S \left| \prod_{\theta' \neq \theta} \left(\ell^{\sigma}_{\theta,\theta'}(s)\right)^{\alpha(\theta')} \right| d\sigma_\theta$$

Assumption 2. *The expectation $M^{\sigma}_{\theta}(\alpha)$ is finite for every θ and every $\alpha : \Theta \to \mathbb{N}$.*

This assumption says that the log-likelihood ratios have finite moments, ruling out for example the signal structure

$$\begin{array}{ccc} & s_1 & s_2 \\ \theta_1 & 0 & 1 \\ \theta_2 & \frac{1}{2} & \frac{1}{2} \end{array}$$

where the signal realization s_1 is perfectly revealing of the state θ_2.

Let $\mathcal{E}$ be the class of all signals satisfying Assumption 2. An *information cost function* is any map $C : \mathcal{E} \to [0, \infty)$. Pomatto, Strack and Tamuz (2020) propose four axioms that such a cost function should further satisfy.

Axiom 1 (Consistency with the Blackwell order)**.** *If σ dominates σ' in the Blackwell order, then $C(\sigma) \geq C(\sigma')$.*

That is, more informative signals are more costly to acquire.

Definition 5.12 (Combining Independent Signals). *For any two signals $\sigma : \Theta \to \Delta(S)$ and $\sigma' : \Theta \to \Delta(S')$, let $\sigma \otimes \sigma'$ denote the product signal*

$$\sigma \otimes \sigma' : \Theta \to \Delta(S \times S')$$

where $\sigma \otimes \sigma'(s, s' \mid \theta) = \sigma(s \mid \theta)\sigma(s' \mid \theta)$.

Axiom 2 (Additivity with respect to independent experiments)**.** *For any two signals σ and σ', $C(\sigma \otimes \sigma') = C(\sigma) + C(\sigma')$.*

That is, the cost of acquiring two (conditionally) independent signals is equal to the sum of their costs. This axiom imposes a constant marginal cost on information similar to the one used to motivate Example 5.14.

Definition 5.13 (Diluting Signals). *For any signal σ, the α-*dilution *of σ, denoted $\alpha \cdot \sigma$, is a signal where with probability α the realization of σ is observed, and otherwise a completely uninformative signal is observed. Formally, $\alpha \cdot \sigma$ is a map from Θ to $S \cup \{\emptyset\}$ where the signal outcome $\emptyset$ has a constant $1 - \alpha$ probability at every state $\theta \in \Theta$, and the remaining probability is assigned to S in proportion to σ.*

Axiom 3 (Linearity in the "dilution" of the experiment)**.** $C(\alpha \cdot \sigma) = \alpha \cdot C(\sigma)$ *for every signal σ and weight $\alpha \in [0, 1]$.*

That is, the cost of a signal is linear in the probability that it generates information.

Remark 5.9. Every posterior separable cost function $C(p, \tau) = \mathbb{E}_{q \sim \tau}[\Phi_p(q)]$ satisfies Axiom 3. To see this, observe that the distribution over posterior beliefs given the diluted signal $\alpha \cdot \sigma$, denoted $\tau_{\alpha \cdot \sigma}$, is the convex combination that puts weight α on the distribution τ_σ generated by σ, and weight $1 - \alpha$ on the prior. So

$$\begin{aligned} C(p, \tau_{\alpha \cdot \sigma}) &= \mathbb{E}_{q \sim \alpha \tau_\sigma + (1-\alpha)\delta_p}[\Phi_p(q)] \\ &= \alpha \mathbb{E}_{q \sim \tau_\sigma}[\Phi_p(q)] + (1 - \alpha)\Phi_p(p) \\ &= \alpha \cdot C(p, \tau_\sigma) \end{aligned}$$

where the second equality uses that C is affine in τ and the third uses that $\Phi_p(p) = 0$ in the definition of a posterior separable cost function.

The final axiom imposes continuity of the cost function with respect to a nonstandard (pseudo)-metric given below.[4]

Definition 5.14. *Given an upper bound $N \geq 1$, define*

$$d_N(\sigma, \sigma') = \max_{\theta \in \Theta} d_{TV}(\sigma_\theta, \sigma'_\theta) + \max_{\theta \in \Theta} \max_{\alpha \in \{0, \ldots, N\}^n} |M^\sigma_\theta(\alpha) - M^{\sigma'}_\theta(\alpha)|$$

where d_{TV} denotes the total variation distance.

Two signals σ and σ' are close under this pseudo-metric if for every state θ, the induced distributions of log-likelihood ratios are close in total-variation distance and additionally have similar moments, for any vector of moments lower or equal to $(N, \ldots, N)$.

Axiom 4 (Continuity.)**.** *The function C is uniformly continuous with respect to d_N.*

[4] This is a pseudometric rather than a metric, since $d_N(\sigma, \sigma')$ is equal to zero for $\sigma \neq \sigma'$ if they induce the same distribution over posterior beliefs.

Remark 5.10. The topology of weak convergence of likelihood ratios and the topology of convergence of likelihood ratios in total variation distance are both more standard. But no cost function which satisfies Axioms 1-3 is continuous in these alternative topologies. To see this, let θ be the unknown bias of a coin, and let σ_n be the signal where with probability $1/n$ the outcome of n independent flips of this coin is observed, and otherwise no information is revealed. Axioms 1-3 imply that $C(\sigma_n) = C(\sigma_{n'})$ for all finite n, n'. But the likelihood ratios of these signals converge in the weak topology (and in the total variation topology) to those of the signal that produces no information, and thus a stronger form of Axiom 4 based on either of these alternative topologies would require these signals to all have zero cost.

Proposition 16. *The cost function $C : \mathcal{E} \to \mathbb{R}$ satisfies Axioms 1-4 if and only if there exists a unique collection of $\mathbb{R}_+$-valued parameters $(\beta_{\theta,\theta'})_{\theta,\theta'\in\Theta}$ such that*

$$C(\sigma) = \sum_{\theta,\theta'\in\Theta} \beta_{\theta,\theta'} \times \underbrace{\int_S \ln \frac{d\sigma_\theta(s)}{d\sigma_{\theta'}(s)} d\sigma_\theta(s)}_{\textit{KL-divergence from } \sigma(\cdot \mid \theta') \textit{ to } \sigma(\cdot \mid \theta)} \tag{5.11}$$

As discussed in Section 5.1.2, the KL-divergence from $\sigma(\cdot \mid \theta')$ to $\sigma(\cdot \mid \theta)$ is a measure of how different the distributions are. The larger this divergence is, the easier it is to reject the hypothesis that the state is θ' when it truly is θ.

Remark 5.11. Axiom 4 can be dispensed with if $\Theta = \{\theta_0, \theta_1\}$, in which case Proposition 16 simplifies to the statement that C satisfies Axioms 1-3 if and only if there exist parameters $\beta_{01}, \beta_{10} \geq 0$ such that

$$C(\sigma) = \beta_{01} D(\sigma(\cdot \mid \theta_0) \| \sigma(\cdot \mid \theta_1)) + \beta_{10} D(\sigma(\cdot \mid \theta_1) \| \sigma(\cdot \mid \theta_0)).$$

A notable contrast with entropy cost is that this cost function permits differentiation between states.

Example 5.15 (Pomatto, Strack and Tamuz (2020)). Suppose the unknown state θ is the US GDP per capita, and the agent holds a uniform prior over $\Theta = \{20,000, \dots, 80,000\}$. Then under entropy cost C_{Ent}, it is equally costly to acquire the signal that reveals whether θ is above or below \$50,000, or the signal that reveals whether θ is even or odd.

The free parameters $\beta_{\theta,\theta'}$ in the representation in (5.11) reflect potentially different costs to distinguishing between different pairs of states. Specifically, we can interpret each $\beta_{\theta,\theta'}$ as the marginal cost of increasing the expected log-likelihood ratio of a signal with respect to states θ and θ' (when θ is the true state). Thus in Example 5.15, we may specify (for example) that it is easier to distinguish between states that are far apart than those that are nearby, i.e., if GDP is in fact 80,000 then it is easier to rule out that GDP is 20,000 than it is to rule out that it is 79,999. In the special case where no pair of states is a priori harder to distinguish than another, then all coefficients are equal to one another.

Example 5.16. Returning to the setting of Example 5.2, where we now use $C(\sigma_\varepsilon^2)$ to mean the cost of acquiring the signal $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, we have

$$C(\sigma_\varepsilon^2) = \sum_{\theta,\theta' \in \Theta} \beta_{\theta,\theta'} \frac{(\theta - \theta')^2}{2\sigma_\varepsilon^2}.$$

This nests the precision of the signal $(1/\sigma_\varepsilon^2)$ as a special case when $\beta_{\theta,\theta'} = \frac{1}{(\theta-\theta')^2}$, with the interpretation that states that are closer (in squared distance) are harder to distinguish.

Remark 5.12. The class of cost functions identified in Proposition 16 does not presuppose that the agent is Bayesian and has a prior belief over the state space. But if the agent does have a prior p, then the cost of the signal that induces distribution τ over posterior beliefs can be restated as

$$\mathbb{E}_{q \sim \tau}[\Phi_p(q)] \tag{5.12}$$

where

$$\Phi_p(q) = \Phi(p) - \sum_{\theta,\theta'} \beta_{\theta,\theta'} \frac{q_\theta}{p_\theta} \ln \left(\frac{q_\theta}{q_{\theta'}} \right) \tag{5.13}$$

so this family of cost functions belongs to the class of posterior-separable cost functions (Definition 5.7), although not to the class of uniform posterior separable cost functions (Definition 5.4).[5]

Exercise 5.5 (G). *Verify that (5.12) is equivalent to the original representation in (5.11) when Φ is defined according to (5.13).*

Hint 2. *Recall from Section 2.2 that the prior p and posterior q at signal realization s are related by* $\log\left(\frac{q(\theta)}{q(\theta')}\right) = \log\left(\frac{p(\theta)}{p(\theta')}\right) + \log\left(\frac{d\sigma_\theta}{d\sigma_{\theta'}}(s)\right)$.

5.4 Additional Exercises

Exercise 5.6 (G). *Suppose $p, q \in \Delta(\mathcal{X} \times \mathcal{Y})$ with p_X and q_X denoting the marginal distributions on $\mathcal{X}$, and $p_{Y|X}$ and $q_{Y|X}$ denoting the respective conditional distributions. Prove that*

$$D(p \| q) = D(p_X \| q_X) + D(p_{Y|X} \| q_{Y|X}).$$

This is known as the chain rule for KL divergence.

Exercise 5.7 (G). *Prove that the entropy cost function in Definiton 5.3 fails Pomatto, Strack and Tamuz (2020)'s Axiom 2.*

[5] Pomatto, Strack and Tamuz (2020) show that a generalization of the representation in (5.11), which permits the parameters $\beta_{\theta,\theta'}$ to depend on the prior, can accommodate uniformly posterior separable cost functions.

Part 2

Learning

Chapter 6

Learning

We now extend the Bayesian framework described in 2.1 to accommodate learning from a sequence of signals. Section 6.3 asks whether an agent will eventually learn the state. Section 6.4 asks whether agents with different prior beliefs will eventually hold similar beliefs. Section 6.5 asks whether agents with different priors expect their disagreement to reduce given information (thus studying a second-order belief). Section 6.6 asks whether agents will commonly learn, i.e., whether agents will eventually believe that other agents believe that they ... have learned the state.

6.1 Preliminaries

Let (Θ, d_Θ) be a complete separable metric space endowed with its Borel σ-algebra Σ, and let $p \in \Delta(\Theta)$ be a (Σ-measurable) probability measure on Θ. As before, we interpret $\theta \sim p$ as an unknown parameter of interest.

The space of signal realizations $(\mathcal{X}, d_X)$ is again a complete separable metric space endowed with its Borel σ-algebra $\mathcal{B}$. There is an infinite sequence of signal realizations $X_1, X_2, \ldots$ taking values in the set $\mathcal{X}^\infty = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots$ where each $\mathcal{X}_t$ is a copy of $\mathcal{X}$. Conditional on the realized θ, signals $X_1, X_2, \ldots$ are generated iid according to a conditional density f_θ, and we refer to each X_t as the period-t signal.

The full state space is $\Omega = \Theta \times \mathcal{X}^\infty = \Theta \times \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots$ and it is equipped with the product σ-algebra $\Sigma \times \mathcal{B}_1 \times \mathcal{B}_2 \times \ldots$ where each $\mathcal{B}_t$ is a copy of $\mathcal{B}$. Throughout, we use P to denote the measure on Ω induced by p and the family $(f_\theta)_{\theta \in \Theta}$, and we use P_θ to denote the conditional measure on $\mathcal{X}^\infty$ when the parameter is θ.

6.2 Binary Example

First consider a single-agent environment with two possible parameter values $\theta \in \{A, B\}$. Each period $t \in \mathbb{Z}_+$ a signal realization from $\{a, b\}$ is generated iid according to

$$\begin{array}{ccc} & a & b \\ A & q & 1-q \\ B & 1-q & q \end{array}$$

where $q > 1/2$. Will an agent who holds a prior belief that the probability of A is $p \in (0,1)$ eventually learn the value of the parameter?

Suppose first that the parameter is $\theta = A$, in which case signals are drawn iid according to $f_A = (q, 1-q)$. For any infinite sequence $\mathbf{x} \in \{a,b\}^\infty$ and any $t \in \mathbb{Z}_+$, let

$$n_t(\mathbf{x}) \equiv \#\{1 \le t' \le t : x_{t'} = a\}$$

denote the number of a-realizations among the first t realizations of $\mathbf{x}$. By the strong law of large numbers, there is a set $\mathcal{X}_0^\infty \subseteq \mathcal{X}^\infty$ of P_A-measure 1 such that

$$\lim_{t\to\infty} \frac{n_t(\mathbf{x})}{t} = q \quad \forall \mathbf{x} \in \mathcal{X}_0^\infty.$$

That is, the limiting fraction of a-realizations is q along each sequence in $\mathcal{X}_0^\infty$.

Since signals are assumed to be conditionally independent, the agent's posterior belief about A following any sequence $(x_1, \ldots, x_t)$ depends only on the count of a and b-realizations. Let n denote the number of a-realizations. Then applying Bayes' rule (Section 2.2.1), the agent's posterior belief is

$$\begin{aligned} P(\theta = A \mid x_1, \ldots, x_t) &= \frac{pq^n(1-q)^{t-n}}{pq^n(1-q)^{t-n} + (1-p)(1-q)^n q^{t-n}} \\ &= \frac{1}{1 + \frac{1-p}{p}\left(\frac{1-q}{q}\right)^{2n-t}} \end{aligned} \tag{6.1}$$

Along any $\mathbf{x} \in \mathcal{X}_0^\infty$ we have

$$\lim_{t\to\infty} P(\theta = A \mid x_1, \ldots, x_t) = \lim_{t\to\infty} \left(1 + \frac{1-p}{p}\left[\left(\frac{1-q}{q}\right)^{2\frac{n_t(\mathbf{x})}{t}-1}\right]^t\right)^{-1} = 1$$

recalling that $q > 1/2$ by assumption.

So the agent's posterior belief P_A-almost surely converges to certainty of the correct value of the parameter, A. An identical argument shows that when the parameter is B then the agent's posterior belief P_B-almost surely converges to certainty of B. Thus the agent (eventually) learns the parameter.

6.3 Doob's Consistency Theorem

A classic result due to Doob (1949) generalizes the individual learning result from the previous section.[1]

[1] Our presentation of this material follows Miller (2018).

Assumption 3 (Identifiability). *If $\theta \neq \theta'$, then $P_\theta \neq P_{\theta'}$.*

In words, Assumption 3 is satisfied if no pair of parameter values induce the same distribution over signals, meaning the parameter is identifiable from its observable implications.

Proposition 17. *Suppose Assumption 3 is satisfied, and let $g : \Theta \to \mathbb{R}$ be any measurable function satisfying $\mathbb{E}|g(\theta)| < \infty$. Then*

$$\lim_{t\to\infty} \mathbb{E}(g(\theta) \mid X_1, X_2, \ldots, X_t) = g(\theta) \quad P\text{-a.s.}$$

In the special case where $g(\theta) = \theta$, the result implies that the posterior expectation of θ converges to its true value almost surely. The following proposition is a Bayesian analogue of the above result, and says that posterior beliefs converge almost surely to a degenerate measure at the true state.

Proposition 18 (Posterior Consistency). *Suppose Assumption 3 holds. Then, there exists a set $\Theta' \subseteq \Theta$ with $p(\Theta') = 1$ such that for every $\theta_0 \in \Theta'$ and every neighborhood B of θ_0,*

$$\lim_{t\to\infty} \mathbb{P}(\theta \in B \mid X_1, X_2, \ldots, X_t) = 1 \quad P_{\theta_0}\text{-a.s.}$$

That is, for any prior distribution, the posterior belief is guaranteed to concentrate in a neighborhood of the true parameter θ—except possibly on a set of parameter values that has measure zero under the agent's prior.

Remark 6.1. The qualification that learning occurs except on a set of "measure zero under the agent's prior" is less harmless than it might initially seem. Consider $\Theta = \mathbb{R}$ where the agent's prior $p \in \Delta(\Theta)$ is a point mass at $\theta = 0$. Then the posterior is also a point mass at zero, so the agent will fail to learn any parameter which is different from 0. But because the set $\mathbb{R}\backslash\{0\}$ has measure zero under the agent's prior, the statement of the result holds in a trivial sense. See also the subsequent discussion in Section 7.2.1.

Remark 6.2. Proposition 18 implies that the agent's posterior belief converges almost surely to a point mass on the true parameter in the topology of weak convergence, i.e., there is a P_θ-measure 1 set of sequences of signal realizations such that

$$d(P^t, \delta_\theta) \to 0$$

along each of these sequences, where d denotes the Levy-Prokhorov metric and $P^t \in \Delta(\Theta)$ denotes the posterior belief after observing the first t coordinates of the sequence. Since d is a metric, we also have that for any alternative prior $\widetilde{p} \in \Delta(\Theta)$ and corresponding posterior belief $\widetilde{P}^t \in \Delta(\Theta)$ (updating to the same t realizations),

$$d(P^t, \widetilde{P}^t) \leq d(P^t, \delta_\theta) + d(\delta_\theta, \widetilde{P}^t).$$

Since the RHS converges to zero almost surely (by Proposition 18), the two agents' posterior beliefs converge to one another almost surely in the topology of weak convergence. The subsequent section provides an even stronger version of this result.

6.4 Merging of Beliefs

Assume that for each $t \geq 1$, a unique conditional probability distribution $P^t(x_1, \ldots, x_t)(C)$ exists for all realized sequences $x_1, \ldots, x_t \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_t$ and unknown events $C \in \mathcal{B}_{t+1} \times \mathcal{B}_{t+2} \times \ldots$.[2] Blackwell and Dubins (1962) show that even if players start out with different prior beliefs, their conditional beliefs will merge to one another in a strong sense.

To state the result formally, recall that for any two probability measures μ_1, μ_2 defined on the same σ-algebra $\mathcal{F}$, *total variation distance* and *absolute continuity* are defined as follows.

Definition 6.1. *The* total variation distance *between μ_1 and μ_2 is*

$$d_{TV}(\mu_1, \mu_2) = \sup_{D \in \mathcal{F}} |\mu_1(D) - \mu_2(D)|$$

Definition 6.2. *If $\mu_2(D) = 0$ implies $\mu_1(D) = 0$ for every $D \in \mathcal{F}$, then μ_1 is* absolutely continuous *with respect to μ_2, denoted $\mu_1 \ll \mu_2$.*

Now we are ready to state the main result:

Proposition 19. *Suppose $p, \widetilde{p} \in \Delta(\Theta)$ are absolutely continuous with respect to one another, and define P, $\widetilde{P}$ to be the measures on Ω induced by the respective priors $p, \tilde{p}$, and the family $(P_\theta)_{\theta \in \Theta}$. Then*

$$\lim_{t \to \infty} d_{TV}(P^t(x_1, \ldots, x_t), \widetilde{P}^t(x_1, \ldots, x_t)) = 0 \quad P\text{-almost surely}$$

That is, if two agents hold different prior beliefs about the parameter but agree on the set of measure-0 events, then their conditional beliefs merge in a strong sense: For *all* measurable future events, agents eventually assign similar probabilities.

Example 6.1. To clarify the difference between this result and the one examined in the previous section, consider the problem of learning the unknown bias of a coin, which is parametrized to $p \in [0, 1]$. A coin whose bias is p lands on Heads with probability p and lands on Tails with probability $1 - p$. Two agents have different prior beliefs on $[0, 1]$ and each observe t independent flips of this coin.

Proposition 18 says that the two agents will eventually learn the bias of the coin as t grows large. Proposition 19 says instead: Suppose the two agents have observed t independent flips of the coin; then, their beliefs over all events regarding the future—e.g., that over half of the remaining coin flips will turn up Heads, or that the limiting fraction of Heads realizations is 1/2—must eventually become close (uniformly across such events).

[2] Blackwell and Dubins (1962) work with the more general notion of "predictive probabilities" P where conditional probabilities can be defined.

6.5 (Expected) Disagreement

We now turn to the impact of information on agents' second-order beliefs—i.e., what they think about what others think. Kartik, Lee and Suen (2021) show that when signals satisfy an MLRP condition, then agents with different beliefs expect information to reduce the extent of disagreement.

Here we assume the set of parameters $\Theta \subseteq \mathbb{R}$ is finite and ordered. Two signals X and $\widetilde{X}$ respectively take values in $\mathcal{X}$ and $\widetilde{\mathcal{X}}$, and we assume that X is Blackwell more informative than $\widetilde{X}$. There are two agents, Ann and Bob, who have common knowledge of the conditional distributions $\{f_{X|\theta}(x \mid \theta)\}_{\theta\in\Theta}$ and $\{f_{\widetilde{X}|\theta}(\widetilde{x} \mid \theta)\}_{\theta\in\Theta}$. But Ann and Bob hold different prior beliefs $f_\theta^A, f_\theta^B \in \Delta(\Theta)$ about the parameter. We use F^A and F^B to denote their perceived joint distributions of $(\theta, X, \widetilde{X})$ (induced by the respective priors and the common knowledge signal distributions), and $\mathbb{E}_A$ and $\mathbb{E}_B$ to denote expectations with respect to these distributions.

Assumption 4. *There is an order $\succ$ on $\mathcal{X}$ and an order $\widetilde{\succ}$ on $\widetilde{\mathcal{X}}$ such that the families $\{f_{X|\theta}(\cdot \mid \theta)\}_{\theta\in\Theta}$ and $\{f_{\widetilde{X}|\theta}(\cdot \mid \theta)\}_{\theta\in\Theta}$ each have MLRP (see Definition 3.2).*

Assumption 5. *Bob's prior f_θ^B likelihood-ratio dominates Ann's prior f_θ^A (see Definition 3.1).*

The agents' prior expectations of the parameter are $\mu_A \equiv \mathbb{E}_A(\theta)$ and $\mu_B \equiv \mathbb{E}_B(\theta)$. We are interested in Ann's prior expectation of Bob's posterior expectation (updated to X), and Bob's prior expectation of Ann's posterior expectation (updated to X), respectively denoted by

$$\mu_{AB}(X) \equiv \mathbb{E}_A[\mathbb{E}_B(\theta \mid X)]$$
$$\mu_{BA}(X) \equiv \mathbb{E}_B[\mathbb{E}_A(\theta \mid X)]$$

Proposition 20. *Suppose Assumptions 4 and 5 are satisfied. If X is Blackwell more informative than $\widetilde{X}$, then*

$$\mu_A \leq \mu_{AB}(X) \leq \mu_{AB}(\widetilde{X}) \leq \mu_B$$

$$\mu_A \leq \mu_{BA}(\widetilde{X}) \leq \mu_{BA}(X) \leq \mu_B$$

That is, Ann expects that a more informative experiment will, in expectation, bring Bob's posterior mean closer to Ann's prior, and vice versa. These are both subjective statements, and indeed only one of Ann and Bob can be correct.

We'll prove this proposition using the following relationships, which are left as an exercise.

Exercise 6.1 (G). *Prove the following statements:*

(a) $F^B_{\theta|X}(\theta \mid X = x)$ *first-order stochastically dominates* $F^A_{\theta|X}(\theta \mid X = x)$ *for every signal realization* $x \in \mathcal{X}$

(b) $F^B_{X|\widetilde{X}}(X \mid \widetilde{X} = \tilde{x})$ *first-order stochastically dominates* $F^A_{X|\widetilde{X}}(X \mid \tilde{X} = \tilde{x})$ *for every signal realization* $\tilde{x} \in \widetilde{\mathcal{X}}$

Proof. Part (a) of Exercise 6.1 implies $\int \theta dF^A_{\theta|X}(\theta \mid x) \leq \int \theta dF^B_{\theta|X}(\theta \mid x)$ for every realization x, so also

$$\int \int \theta dF^A_{\theta|X}(\theta \mid x) dF^A_X(x) \leq \int \int \theta dF^B_{\theta|X}(\theta \mid x) dF^A_X(x). \tag{6.2}$$

By assumption that $\{f_{X|\theta}(\cdot \mid \theta)\}_{\theta\in\Theta}$ has MLRP, the integral $\int \theta dF^B_{\theta|X}(\theta \mid x)$ is an increasing function of x. Moreover, Part (b) of Exercise 6.1 says that F^B_X first-order stochastically dominates F^A_X (taking $\widetilde{X}$ to be any constant signal). Thus

$$\int \int \theta dF^B_{\theta|X}(\theta \mid x) dF^A_X(x) \leq \int \int \theta dF^B_{\theta|X}(\theta \mid x) dF^B_X(x). \tag{6.3}$$

Together, (6.2) and (6.3) imply

$$\int \int \theta dF^A_{\theta|X}(\theta \mid x) dF^A_X(x) \leq \int \int \theta dF^B_{\theta|X}(\theta \mid x) dF^A_X(x) \leq \int \int \theta dF^B_{\theta|X}(\theta \mid x) dF^B_X(x)$$

which is precisely the desired inequality $\mu_A \leq \mu_{AB}(X) \leq \mu_B$. It follows by identical arguments that $\mu_A \leq \mu_{BA}(X) \leq \mu_B$.

To show that $\mu_{AB}(\widetilde{X}) \geq \mu_{AB}(X)$, we use the fact that (since X Blackwell-dominates $\widetilde{X}$) we can generate the two variables in such a way that $\widetilde{X}$ is conditionally independent of θ conditional on X.[3] Then on this probability space

$$\begin{aligned}
\mu_{AB}(\widetilde{X}) &= \mathbb{E}_A\left[\mathbb{E}_B\left(\theta \mid \widetilde{X}\right)\right] & \\
&= \mathbb{E}_A\left[\mathbb{E}_B\left(\mathbb{E}_B\left(\theta \mid X, \widetilde{X}\right) \mid \widetilde{X}\right)\right] & \text{by L.I.E.} \\
&= \mathbb{E}_A\left[\mathbb{E}_B\left(\mathbb{E}_B\left(\theta \mid X\right) \mid \widetilde{X}\right)\right] & \text{since } \widetilde{X} \perp\!\!\!\perp \theta \mid X \\
&= \int \int \mathbb{E}_B(\theta \mid x) dF^B_{X|\widetilde{X}}(x \mid \tilde{x}) dF_A(\tilde{x}) & \\
&\geq \int \int \mathbb{E}_B(\theta \mid x) dF^A_{X|\widetilde{X}}(x \mid \tilde{x}) dF_A(\tilde{x}) & \\
&= \mathbb{E}_A\left[\mathbb{E}_A\left(\mathbb{E}_B\left(\theta \mid X\right) \mid \widetilde{X}\right)\right] & \\
&= \mathbb{E}_A\left[\mathbb{E}_B\left(\theta \mid X\right)\right] & \text{by L.I.E.} \\
&= \mu_{AB}(X) &
\end{aligned}$$

where the crucial inequality follows by observing that $\mathbb{E}_B(\theta \mid x)$ is an increasing function of x (by Assumption 4) while $F^B_{X|\widetilde{X}}(\cdot \mid \tilde{x})$ first-order stochastically dominates $F^A_{X|\widetilde{X}}(\cdot \mid \tilde{x})$ for every realization of $\tilde{x}$ (by Part (b) of Exercise 6.1).

Since the previous arguments apply to show also that $\mu_{AB}(\widetilde{X}) \leq \mu_B$, we are done. ∎

[3] See Remark 4.1 for further detail. Note also that the correlation between X and $\widetilde{X}$ is irrelevant for the comparison of $\mu_{AB}(X)$ and $\mu_{AB}(\widetilde{X})$.

6.6 Common Learning

Suppose Assumption 3 (Identifiability) holds, so that agents eventually learn the true parameter. Does this imply that agents will eventually have *common knowledge* of the true parameter? Cripps et al. (2008) adapt Monderer and Samet (1989)'s definition of common q-belief for the present learning environment, and show that individual learning does imply common learning when the set of signal realizations is finite, but that this implication may otherwise fail.

In what follows recall that each state $\omega \in \Omega = \Theta \times \mathcal{X}^\infty$ describes both the value of the parameter and the infinite sequence of signal profiles. As before, P_θ denotes the measure on $\mathcal{X}^\infty$ conditional on parameter θ, and again assume that Θ is finite. There are two agents $i = 1, 2$, and (different from the previous sections) we decompose $\mathcal{X} = \mathcal{X}^1 \times \mathcal{X}^2$ where $\mathcal{X}^i$ denotes the set of agent i signal realizations. Each agent privately observes their own signal each period. We use $h_{it}(\omega) = (x_1^i(\omega), \ldots, x_t^i(\omega))$ for agent i's history at time t when ω is the realized state, and $\mathcal{H}_{it}$ to denote the filtration induced by agent i's histories.

Definition 6.3. *For any $q \in [0,1]$ and (measurable) event F, agent i q-*believes *in F at time t on*

$$B_{it}^q(F) = \{\omega \in \Omega \mid P(F \mid h_{it}(\omega)) \geq q\}$$

Definition 6.4. *For any $q \in [0,1]$, there is* common q-belief *in F at time t on*

$$C_t^q(F) = \bigcap_{n \geq 1} [B_t^q]^n(F)$$

where $B_t^q(F) = B_{1t}^q(F) \bigcap B_{2t}^q(F)$.

Definition 6.5 (Individual Learning). *Agent i* learns *θ if for each $q \in (0,1)$ there exists $T < \infty$ such that*

$$P_\theta(B_{it}^q(\{\theta\} \times \mathcal{X}^\infty)) > q \quad \forall t > T$$

Equivalently: $\lim_{t\to\infty} P_\theta(B_{it}^q(\{\theta\} \times \mathcal{X}^\infty)) = 1$ *for all $q \in (0,1)$. Agent i* individually learns *if the agent learns each $\theta \in \Theta$.*

Definition 6.6 (Common Learning). *Agents* commonly learn *θ if for each $q \in (0,1)$ there exists $T < \infty$ such that*

$$P_\theta(C_t^q(\{\theta\} \times \mathcal{X}^\infty)) > q \quad \forall t > T$$

Equivalently: $\lim_{t\to\infty} P_\theta(C_t^q(\{\theta\} \times \mathcal{X}^\infty)) = 1$ *for all $q \in (0,1)$. Agents* commonly learn *if they commonly learn each $\theta \in \Theta$.*

Clearly if signals are perfectly correlated (or public), so that $P(\theta \mid \mathcal{H}_{1t}) = P(\theta \mid \mathcal{H}_{2t})$ for all θ and t, then individual learning implies common learning. This result also holds at the other extreme of independent signals.

Proposition 21. *Suppose agents individually learn, and their signals are conditionally independent given the parameter. That is, there exist families $(P^i_\theta)_{\theta\in\Theta}$, with each $P^i_\theta \in \Delta(\mathcal{X}^i)$, such that $P_\theta(A\times B) = P^1_\theta(A)P^2_\theta(B)$ for each $\theta\in\Theta$ and measurable $A\subseteq \mathcal{X}^1$, $B\subseteq\mathcal{X}^2$. Then, agents commonly learn.*

Cripps et al. (2008) proves this proposition using a result from Monderer and Samet (1989) (adapted to the present learning context).

Lemma 2. *Agents commonly learn if and only if for every $\theta\in\Theta$ and $q\in(0,1)$, there is a sequence of events F_t and a period T such that for all $t>T$,*

(a) $F_t \subseteq B^q_t(\theta)$ ("θ is q-believed on F_t at time t")

(b) $P_\theta(F_t) > q$ ("probability of F_t is sufficiently high")

(c) $F_t\subseteq B^q_{it}(F_t)$ for $i=1,2$ ("F_t is evident q-belief at time t")

We'll now prove Proposition 21.

Proof. Henceforth write $\{\theta\}$ for the event $\{\theta\}\times\mathcal{X}^\infty$. Define $F_t = \{\theta\}\cap B^{\sqrt{q}}_t(\theta)$ to be the set of states at which θ is true and both agents $\sqrt{q}$-believe it. We'll verify that the conditions of Lemma 2 hold for the sequence of events $(F_t)^\infty_{t=1}$, from which Proposition 21 follows.

First observe that

$$\begin{aligned} F_t &\subseteq B^{\sqrt{q}}_t(\theta) && \text{by definition of } F_t \\ &\subseteq B^q_t(\theta) && \text{since } q<\sqrt{q} \end{aligned}$$

yielding Part (a) of Lemma 2. Part (b) holds since individual learning implies that there exists $T<\infty$ such that for both agents $i=1,2$,

$$P_\theta\left(B^{\sqrt{q}}_{it}(\theta)\right) > \sqrt{q} \quad \forall t>T$$

and thus

$$P_\theta(F_t) = P_\theta\left(B^{\sqrt{q}}_{1t}(\theta)\right)P_\theta\left(B^{\sqrt{q}}_{2t}(\theta)\right) > q \quad \forall t>T$$

from the assumption of conditional independence.

It remains to show Part (c). First rewrite the set $B^q_{1t}(F_t)$ as follows:

$$\begin{aligned} B^q_{1t}(F_t) &= \{\omega \mid \mathbb{E}\left[\mathbb{1}_{F_t}\mid\mathcal{H}_{1t}\right]\geq q)\} && \text{by definition of } B^q_{1t} \\ &= \left\{\omega \mid \mathbb{E}\left[\mathbb{1}_{B^{\sqrt{q}}_{1t}(\theta)}\mathbb{1}_{B^{\sqrt{q}}_{2t}(\theta)\cap\{\theta\}}\mid\mathcal{H}_{1t}\right]\geq q\right\} && \text{by definition of } F_t \\ &= \left\{\omega \mid \mathbb{1}_{B^{\sqrt{q}}_{1t}(\theta)}\mathbb{E}\left[\mathbb{1}_{B^{\sqrt{q}}_{2t}(\theta)\cap\{\theta\}}\mid\mathcal{H}_{1t}\right]\geq q\right\} && \text{since } B^{\sqrt{q}}_{1t}(\theta)\in\mathcal{H}_{1t} \\ &= B^{\sqrt{q}}_{1t}(\theta)\cap B^q_{1t}\left(B^{\sqrt{q}}_{2t}(\theta)\cap\{\theta\}\right) \end{aligned}$$

By definition we have that $F_t\subseteq B^{\sqrt{q}}_{1t}(\theta)$. As above, individual learning implies existence of T sufficiently large that $P_\theta\left(B^{\sqrt{q}}_{2t}(\theta)\right) > \sqrt{q}$ for all $t>T$. Since signals are conditionally independent, agent 1's history is uninformative about agent 2's history, implying that

$$P_\theta\left(B_{2t}^{\sqrt{q}}(\theta)\mid\mathcal{H}_{1t}\right)\geq\sqrt{q} \tag{6.4}$$

holds uniformly across agent 1 histories (for all $t > T$). So on F_t (for $t > T$) we have

$$P(B_{2t}^{\sqrt{q}}(\theta)\cap\{\theta\}\mid\mathcal{H}_{1t}) = \underbrace{P_\theta(B_{2t}^{\sqrt{q}}(\theta)\mid\mathcal{H}_{1t})}_{>\sqrt{q}\text{ by (6.4)}}\quad\underbrace{P(\theta\mid\mathcal{H}_{1t})}_{>\sqrt{q}\text{ since }F_t\subseteq B_{1t}^{\sqrt{q}}(\theta)}\quad > q.$$

Apply Lemma 2 and we are done. ■

REMARK 6.3. This proof extends for arbitrary finite numbers of agents, setting $F_t = \{\theta\}\cap B_t^{\sqrt[n]{q}}(\theta)$.

Although common learning is implied by individual learning when agents have either perfect information or no information about the other agent's history, intermediate cases of correlation can break this result.

EXAMPLE 6.2. (Twist on Rubinstein (1989)'s email game.) The unknown parameter is $\theta\in\{\theta',\theta''\}$, where $0\leq\theta'<\theta''\leq 1$. Suppose that every period a signal profile is independently drawn according to:

Probability	Agent-1 Signal	Agent-2 Signal
θ	0	0
$\varepsilon(1-\theta)$	1	0
$(1-\varepsilon)\varepsilon(1-\theta)$	1	1
$(1-\varepsilon)^2\varepsilon(1-\theta)$	2	1
$(1-\varepsilon)^3\varepsilon(1-\theta)$	2	2
$(1-\varepsilon)^4\varepsilon(1-\theta)$	3	2
$(1-\varepsilon)^5\varepsilon(1-\theta)$	3	3
⋮	⋮	⋮

This signal structure generalizes the information structure in the email game from Section 1.3, where $\theta = 1$ corresponds to state a in the email game and $\theta = 0$ corresponds to state b.

Agents observe repeated independent realizations of the signal. Will they commonly learn the game parameter? When θ is restricted to values 0 and 1 (as per Rubinstein (1989)'s email game), the answer is yes.

EXERCISE 6.2 (G). *Prove that common learning occurs if $\theta\in\{\theta',\theta''\}\equiv\{0,1\}$.*

But common learning fails whenever $0 < \theta' < \theta'' < 1$ as agents cannot commonly learn θ'', the parameter placing more weight on the lower signal realizations. Intuitively, when 1 sees the signal k, then he believes with some probability (that can be uniformly lower bounded across histories) that 2 has also observed at least k. And if 2 observes k, then he believes with some probability (that again can be uniformly lower bounded) that 1 observed $k+1$. Since the number of signal realizations is infinite, there is unbounded contagion upwards: The agent always believes with some probability that the other agent believes with some probability that he has observed... such a large signal that he believes that the state is (very likely to be) θ'. And thus we cannot establish common q-belief of θ'' for large q.

The main result in Cripps et al. (2008) establishes that infinite signal realizations are critical to the previous counterexample. When the number of signal realizations is finite, then individual learning always implies common learning.

Assumption 6 (Finite Signal Sets). $|\mathcal{X}^1|, |\mathcal{X}^2| < \infty$

Proposition 22. *If Assumption 6 is satisfied, then individual learning implies common learning.*

A brief idea of the proof follows. Define $\pi^\theta(ij)$ to be the probability of realization $(x_t^1, x_t^2) = (i, j)$ when the parameter is θ, and define

$$\phi^\theta(i) = \sum_{j \in \mathcal{X}^2} \pi^\theta(ij)$$

to be the marginal probability of signal i, with $\phi^\theta \equiv (\phi^\theta(i))_{i \in \mathcal{X}^1}$. Likewise define

$$\psi^\theta(j) = \sum_{i \in \mathcal{X}^1} \pi^\theta(ij)$$

to be the marginal probability of signal j, with $\psi^\theta \equiv (\psi^\theta(j))_{j \in \mathcal{X}^2}$. Then (by the results in Section 6.3), individual learning follows whenever $\phi^\theta \neq \phi^{\theta'}$ and $\psi^\theta \neq \psi^{\theta'}$ for every $\theta \neq \theta'$.

Define $\hat{\phi}_t$ to be the empirical frequency of agent 1 signals and $\hat{\psi}_t$ to be the empirical frequency of agent 2 signals. Under the assumption of individual learning, empirical frequencies must converge to the theoretical frequencies, i.e., for each parameter θ, $\hat{\phi}_t \to \phi^\theta$ and $\hat{\psi}_t \to \phi^\theta$ P_θ-almost surely. Thus each agent eventually assigns a high probability to true θ.

The crucial next step is establishing that when agent 1 assigns a high probability to θ, he believes that agent 2 does as well (and vice versa). To see why this might be the case, let M_1^θ be the $|\mathcal{X}^1| \times |\mathcal{X}^2|$ matrix whose (i,j)-th entry is $\frac{\pi^\theta(ij)}{\phi^\theta(i)}$, i.e. the conditional probability (under θ) that agent 2 observes j given that agent 1 observed i, and define M_2^θ analogously. Then $\hat{\phi}_t M_1^\theta$ is agent 1's expectation of agent 2's realized frequencies (conditional on θ), and $\hat{\phi}_t M_1^\theta M_2^\theta$ is agent 1′s expectation of agent 2′s expectation of agent 1′s realized frequencies (again conditional on θ). Observe (by algebra) that

$$\phi^\theta M_1^\theta = \psi^\theta$$
$$\psi^\theta M_2^\theta = \phi^\theta$$

so $\phi^\theta M_1^\theta M_2^\theta = \phi^\theta$. Indeed the matrix $M_{12}^\theta \equiv M_1^\theta M_2^\theta$ is a Markov transition matrix on $\mathcal{X}^1$ with stationary distribution ϕ^θ, and it is moreover a contraction mapping on $\Delta(\mathcal{X}^1)$. These properties together imply that the higher order beliefs cannot run away from the agent′s first-order belief as they did in Example 6.2.

6.7 Additional Exercises

Exercise 6.3 (G*). *Let $\theta \sim \mathcal{N}(0,1)$ be an unknown parameter. Each agent $i = 1, 2$ observes n signals $X_1^i, \ldots, X_n^i$ where each*

$$X_m^i = \theta + \varepsilon_m^i$$

with $\varepsilon_m^i \sim \mathcal{N}(0,1)$ independent of θ, independent across agents, and independent across signals. Suppose that the true value of θ is strictly positive, and let E_p be the event that the two agents have common p-belief that θ is positive, where $p > 1/2$. What is the probability of E_p under the actual data-generating process?

Chapter 7

Model Uncertainty and Misspecification

We have so far assumed that agents' model of the world is *correctly specified*: Their prior belief over Θ assigns positive probability to the true parameter θ and they update to information correctly, i.e. with knowledge of the true signal generating distribution $(P_\theta)_{\theta\in\Theta}$. Some reasons to question this model of learning include:

- We see substantial and persistent disagreement between individuals, but Sections 6.3 and 6.4 imply that agents will eventually hold similar beliefs.
- It is unclear how agents came to know $(P_\theta)_{\theta\in\Theta}$.
- The assumption that agents perceive only one signal-generating distribution $(P_\theta)_{\theta\in\Theta}$ as possible means that agents never abandon their model, even as evidence accumulates against it. As we discuss in Section 7.1.1, this dogmatism has some strange implications.

This section relaxes the standard learning model by allowing for *model uncertainty* (Section 7.1) and *model misspecification* (Section 7.2). In the former class of models, agents hold non-degenerate beliefs over the signal generating distribution. In the second, agents assign probability zero to the true parameter.

7.1 Model Uncertainty

7.1.1 Motivation

Recall the binary setting from Section 6.2: There is an unknown parameter $\theta \in \{A, B\}$, and each period $t \in \mathbb{Z}_+$ a signal is generated iid according to

$$\begin{array}{ccc} & a & b \\ A & q & 1-q \\ B & 1-q & q \end{array}$$

where $q > 1/2$. Agents may hold different (non-degenerate) prior beliefs $\pi_i \in \Delta(\Theta)$ about the parameter, but the value of q is common knowledge.

In Section 6.2, we observed that these agents almost surely learn the true parameter as the sample size grows large, and moreover their disagreement about the parameter vanishes. This is because (1) agents assign probability 1 to the event in which the limiting fraction of a-realizations is either q or $1-q$, and (2) the parameter is identified, so for either of these limiting frequencies agents (eventually) assign probability 1 to the correct parameter value.

What happens along sequences in which the limiting frequency is neither $(q, 1-q)$ nor $(1-q, q)$? Although agents assign probability zero to this event, sampling variation can explain any empirical frequency of a and b realizations (however surprising) in finite sequences. Thus Bayes' rule yields well-defined posterior beliefs.

For example, suppose $q \in (1/2, 1)$ and let $\mathbf{x}$ be the (infinite) sequence of a-realizations. For any t, the unconditional probability of the event that all t realizations are a is

$$\pi^i_A \cdot q^t + (1-\pi^i_A) \cdot (1-q)^t$$

where π^i_A denotes the prior probability of A. This expression converges to zero as t grows large but is strictly positive for every t. The agent's limiting belief along $\mathbf{x}$ can thus be computed to be

$$\lim_{t\to\infty} P^i(\theta = A \mid \mathbf{x}_t) = \lim_{t\to\infty} \frac{1}{1 + \frac{1-\pi^i_A}{\pi^i_A}\left(\frac{1-q}{q}\right)^t} = 1$$

So the agent is increasingly convinced that the state is A, even as the observed sequence grows increasingly unlikely under the agent's model. Even more striking, as signals accumulate in the frequency $(1, 0)$, the agent becomes increasingly confident that future signals will appear in the frequency $(q, 1-q)$! These conclusions are a consequence of the agent's dogmatic view of the signal generating distribution—he is unwilling to abandon this model even as mounting evidence points to its error.

7.1.2 Expanded Framework

We can introduce *model uncertainty* into this learning model by expanding the state space to $\Omega = \Theta \times \Gamma \times \mathcal{X}^\infty$ where the new parameter γ indexes the signal-generating distribution, and the parameters θ and γ jointly determine a family $(P_{\theta,\gamma})_{\theta\in\Theta,\gamma\in\Gamma}$ of conditional distributions over signals. The key distinction between θ and γ is that only θ is payoff-relevant. We'll use P^i to denote agent i's subjective prior belief on Ω, which is common knowledge to all agents.

If people do not in fact have dogmatic beliefs about the signal-generating distribution, a natural question is whether modeling agents in this way is still a good abstraction, in the sense that the qualitative insights of this model are robust to introduction of a small amount of model uncertainty. Acemoglu, Chernozhukov and Yildiz (2015) demonstrate one important sense in which this is not so.

7.1.3 Failure of Asymptotic Agreement

For any infinite sequence $\mathbf{x} \in \mathcal{X}^\infty$, write

$$\phi^i_{\theta,t} \equiv P^i(\theta \mid x_1, \dots x_t)$$

for the posterior probability that agent i assigns to θ following the first t realizations of the sequence $\mathbf{x}$. Further define

$$\phi^i_{\theta,\infty}(\mathbf{x}) = \lim_{t\to\infty} \phi^i_{\theta,t}(\mathbf{x}) \tag{7.1}$$

to be the asymptotic posterior probability that agent i assigns to θ along sequence $\mathbf{x}$.

Definition 7.1. *Say that* asymptotic agreement *occurs if for each agent i,*

$$P^i(\phi^1_{\theta,\infty} = \phi^2_{\theta,\infty}) = 1 \quad \forall \theta \in \Theta$$

That is, both agents believe their asymptotic beliefs will be identical.

When agents hold a dogmatic belief about the signal-generating distribution, asymptotic agreement occurs whenever the parameter is identified (Proposition 18). But Acemoglu, Chernozhukov and Yildiz (2015) show that asymptotic agreement can fail when an arbitrarily small amount of model uncertainty is introduced. The basic idea behind this fragility can be seen through this following example from their paper.

Let $\Theta = \{A, B\}$, with each agent i's prior about the parameter denoted by $\pi^i \equiv (\pi^i_A, \pi^i_B)$. Agent i believes that signals are generated iid from the set $\{a, b\}$ with state-dependent distribution

$$\begin{array}{ccc} & a & b \\ A & \gamma & 1-\gamma \\ B & 1-\gamma & \gamma \end{array}$$

where γ is unknown and distributed according to G^i with density

$$g^i(\gamma) = \begin{cases} \varepsilon + \frac{1-\varepsilon}{\lambda} & \text{if } \gamma \in (\gamma^i - \lambda/2, \gamma^i + \lambda/2) \\ \varepsilon & \text{otherwise} \end{cases}$$

for some $\gamma^i > 1/2$. Assume that γ^1 and γ^2 are different from one another. This density is depicted in Figure 7.1.

The limit as $\varepsilon \to 0$ and $\lambda \to 0$ returns the model in which each agent i dogmatically believes the signal structure to be given by

$$\begin{array}{ccc} & a & b \\ A & \gamma^i & 1-\gamma^i \\ B & 1-\gamma^i & \gamma^i \end{array}$$

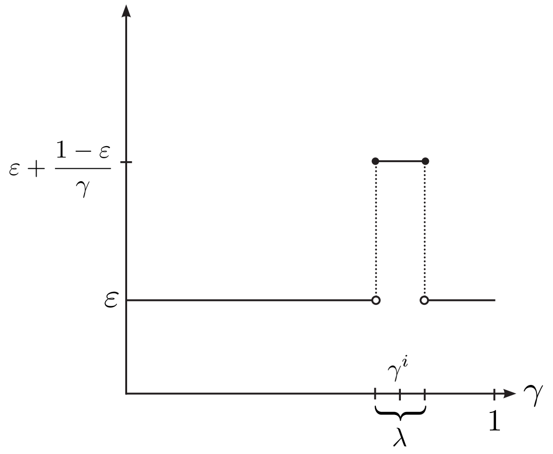

Figure 7.1: Depiction of g^i.

At this limit, asymptotic agreement holds.

Now suppose ε and λ are strictly positive and λ is small (specifically, let $\lambda < |\gamma^1 - \gamma^2|$ and suppose $\gamma^i - \frac{\lambda}{2} > \frac{1}{2}$ for each agent i). As in Section 6.2, define

$$n_t(\mathbf{x}) \equiv \#\{1 \leq t' \leq t : \mathbf{x}_{t'} = a\} \quad \forall \mathbf{x} \in \mathcal{X}^\infty$$

to be the count of a-realizations among the first t realizations of $\mathbf{x}$, and let

$$\rho(\mathbf{x}) = \lim_{t\to\infty} n_t(\mathbf{x})/t \quad \forall \mathbf{x} \in \mathcal{X}^\infty$$

be the asymptotic frequency of a-realizations along $\mathbf{x}$.

The following lemma provides a simple expression for the agent's asymptotic belief (7.1) on the set of sequences $\widetilde{\mathcal{X}}^\infty \subseteq \mathcal{X}^\infty$ where the limiting frequency $\rho(\mathbf{x})$ exists.

Lemma 3 (Acemoglu, Chernozhukov and Yildiz (2015)). *For every sequence* $\mathbf{x} \in \widetilde{\mathcal{X}}^\infty$,

$$\phi^i_{A,\infty}(\mathbf{x}) = \left(1 + \frac{1-\pi^i_A}{\pi^i_A} \cdot \frac{f^i_B(\rho(\mathbf{x}), 1-\rho(\mathbf{x}))}{f^i_A(\rho(\mathbf{x}), 1-\rho(\mathbf{x}))}\right)^{-1}$$

where $\frac{f^i_B(\rho(\mathbf{x}),1-\rho(\mathbf{x}))}{f^i_A(\rho(\mathbf{x}),1-\rho(\mathbf{x}))}$ *is the asymptotic likelihood ratio under agent i's subjective model.*

In the running example of this section, the asymptotic likelihood ratio can be simplified to

$$\frac{f^i_B(\rho, 1-\rho)}{f^i_A(\rho, 1-\rho)} = \frac{g^i(1-\rho)}{g^i(\rho)}$$

This ratio takes on either of three possible values. For any $\rho \in (\gamma^i - \lambda/2, \gamma^i + \lambda/2)$,

$$\frac{g^i(1-\rho)}{g^i(\rho)} = \frac{\varepsilon\lambda}{1-\varepsilon(1-\lambda)}$$

which converges to zero as ε and λ grow small (implying $\phi^i_{A,\infty} \to 1$). By a mirror argument, if the limiting frequency of a-realizations is some $\rho \in (1 - \gamma^i - \lambda/2, 1 - \gamma^i + \lambda/2)$, then

$$\frac{g^i(1-\rho)}{g^i(\rho)} = \frac{1-\varepsilon(1-\lambda)}{\varepsilon\lambda}$$

which converges to ∞ as ε and λ grow small (implying $\phi_{A,\infty} \to 0$). For all other limiting frequencies, the asymptotic likelihood ratio is simply $\frac{g^i(1-\rho)}{g^i(\rho)} = 1$. These unlikely signal sequences are considered possible but uninformative about the parameter.

Applying Lemma 3, Figure 7.2 depicts agent i's asymptotic posterior as a function of the limiting signal frequency.

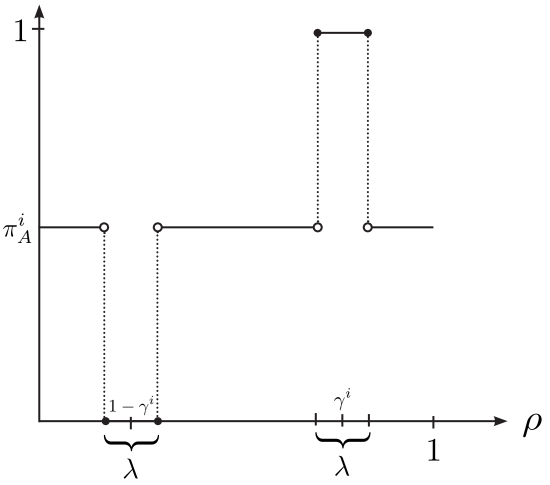

Figure 7.2: Agent i's asymptotic posterior in the limit as $\varepsilon \to 0$.

In the limit as $\varepsilon \to 0$ and $\lambda \to 0$, each agent i is increasingly sure that the limiting frequency ρ will either be close to γ^i or $1-\gamma^i$, so he believes that he will (approximately) learn the parameter. But when a sequence of signals has a long-run frequency that leads agent 1 to learn $\theta = A$ or $\theta = B$, agent 1 knows that this sequence has led agent 2 to consider the signal uninformative, in which case agent 2's limiting belief is the same as his prior. Likewise whenever agent 2 believes the signal sequence to be informative about θ, he knows that agent 1 considers the signal sequence to be uninformative. So not only does asymptotic agreement fail, but we have the stronger conclusion that the limiting beliefs ϕ^1_∞ and ϕ^2_∞ are different on *all* sample paths. Figure 7.3 depicts $|\phi^1_{A,\infty} - \phi^2_{A,\infty}|$ as a function of the limiting signal frequency.

To summarize, asymptotic agreement holds in the limiting model $\varepsilon = 0, \lambda = 0$ (with no model uncertainty), but fails when the model is perturbed to include an arbitrarily small amount of model uncertainty via $\varepsilon > 0, \lambda > 0$.

Remark 7.1. As in Section 6, there is no ground truth—whether asymptotic agreement does or doesn't hold is determined solely with respect to the agents' subjective beliefs.

Remark 7.2. In this example, the two agents' prior beliefs on $\Theta \times \Gamma$ are absolutely continuous with respect to one another. So Proposition 19 tells us that their beliefs about future signal realizations will eventually merge. But (θ, γ) is not identified: For example, $(A, 1)$ and $(B, 0)$ identically lead to a degenerate distribution on the infinite sequence of a-realizations. Thus asymptotic agreement about the expanded parameter (θ, γ) is not guaranteed from the results of Sections 6.3 and 6.4.

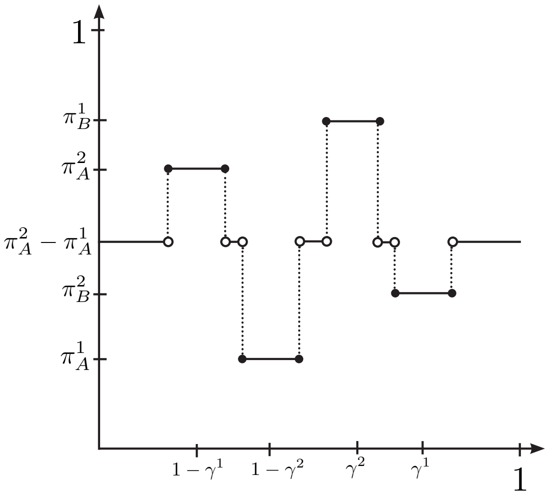

Figure 7.3: Asymptotic disagreement $|\phi^1_{A,\infty} - \phi^2_{A,\infty}|$ in the limit as $\varepsilon \to 0$, for parameter values $\pi_B^1 > \pi_A^2 > \pi_B^2 > \pi_A^1$.

7.2 Misspecified Learning

Next suppose the agent is not simply uncertain about the signal-generating distribution, but in fact rules out the true distribution.

Example 7.1. Let $\Theta = \{A, B, C\}$ where the conditional distributions over signal realizations $\{a, b\}$ are given as follows:

	a	b
A	4/5	1/5
B	1/2	1/2
C	2/3	1/3

The agent has a uniform prior on $\{A, B\}$, but the true parameter is C. Given repeated independent observations from the distribution $(2/3, 1/3)$, will the agent's beliefs converge and if so to what limiting belief?

7.2.1 Role of KL Divergence

Intuitively, we may expect that the agent's beliefs converge to certainty of the parameter whose distribution is "closer" to the true distribution. The right notion of closeness here turns out to be KL Divergence (Section 5.1.2).

Here is a heuristic argument for how KL divergence emerges. Suppose the agent only considers parameter values $\theta = A$ and $\theta = B$ to be possible, where the prior probability of $\theta = A$ is $\pi \in (0, 1)$. We'll use $f_\theta(x)$ to denote the conditional probability of signal realization x when the parameter is θ. The agent observes a sequence of signals drawn iid according to f_{θ^*}, where the "true" parameter value θ^* may be different from both A and B.

For any signal sequence $\mathbf{x}_t = (x_1, \ldots, x_t)$, the conditional probability of A can be rewritten

$$\begin{aligned}
&\mathbb{P}(\theta = A \mid \mathbf{x}_t) \\
&\quad = \left(1 + \frac{1-\pi}{\pi}\left(\prod_{i=1}^{t} \frac{f_B(x_i)}{f_A(x_i)}\right)\right)^{-1} \\
&\quad = \left(1 + \frac{1-\pi}{\pi}\left(\prod_{i=1}^{t} \frac{f_B(x_i)/f_{\theta^*}(x_i)}{f_A(x_i)/f_{\theta^*}(x_i)}\right)\right)^{-1} \\
&\quad = \left(1 + \frac{1-\pi}{\pi}\exp\left(-\log\left(\prod_{i=1}^{t} \frac{f_{\theta^*}(x_i)}{f_B(x_i)}\right) + \log\left(\prod_{i=1}^{t} \frac{f_{\theta^*}(x_i)}{f_A(x_i)}\right)\right)\right)^{-1} \\
&\quad = \left(1 + \frac{1-\pi}{\pi}\exp\left(-n \cdot \left(\frac{1}{t}\sum_{i=1}^{t} \log\left(\frac{f_{\theta^*}(x_i)}{f_B(x_i)}\right) - \frac{1}{t}\sum_{i=1}^{t} \log\left(\frac{f_{\theta^*}(x_i)}{f_A(x_i)}\right)\right)\right)\right)^{-1}
\end{aligned}$$

and for large t this final display is approximately equal to

$$\left(1 + \frac{1-\pi}{\pi}\exp\left(-t \cdot \left(D(f_{\theta^*} \| f_B) - D(f_{\theta^*} \| f_A)\right)\right)\right)^{-1} \tag{7.2}$$

If $\theta^* \in \{A, B\}$, then either $D(f_{\theta^*} \| f_A) = 0 < D(f_{\theta^*} \| f_B)$ (in which case the expression in (7.2) converges to 1) or $D(f_{\theta^*} \| f_B) = 0 < D(f_{\theta^*} \| f_A)$ (in which case the expression in (7.2) converges to 0). In either case beliefs converge to certainty of the true parameter, as previously implied by Proposition 18 (Section 6.3).

Suppose now that $\theta^* \notin \{A, B\}$. Proposition 18 no longer applies: Doob (1949)'s consistency result is with respect to a P-measure 1 set of sequences, (where P is the agent's prior on $\Theta \times \mathcal{X}^\infty$), but in this example θ^* falls in the P-measure zero set on which consistency is not guaranteed. Indeed, in Section 6.3 we made no reference to a "true" distribution—consistency was demonstrated within the agent's subjective model.

But (7.2) is useful even when θ^* has zero probability under the agent's prior. Specifically, when $D(f_{\theta^*}\|f_A) < D(f_{\theta^*}\|f_B)$, then (7.2) converges to 1 as $t \to \infty$, yielding certainty of $\theta = A$, and when $D(f_{\theta^*}\|f_A) > D(f_{\theta^*}\|f_B)$, then (7.2) converges to zero as $t \to \infty$. So the agent's beliefs concentrate on the parameter that induces a distribution over signals that is closest in Kullback-Liebler divergence to the true distribution.

Berk (1966) establishes this result more generally. We'll use the notation of Section 6.1, introducing θ^* as new notation for the true parameter, and assuming that the observed signals are drawn iid according to the density f_{θ^*} (with P_{θ^*} denoting the induced measure on $\mathcal{X}^\infty$). To simplify exposition, assume that Θ is finite.

Proposition 23 (Berk (1966))**.** *Let*

$$A \equiv \operatorname*{arg\,min}_{\theta \in Supp(P)} D(f_{\theta^*}\|f_\theta)$$

be the set of parameters in the support of the agent's prior that minimize KL divergence to the true distribution. Then

$$\lim_{t\to\infty} P(A \mid X_1, \ldots, X_t) = 1 \quad P_{\theta^*}\text{-a.s.}$$

Example 7.2. Returning to Example 7.1, since

$$D(f_C\|f_A) = (2/3)\cdot\log\left(\frac{2/3}{4/5}\right) + (1/3)\cdot\log\left(\frac{1/3}{1/5}\right) \approx 0.021$$
$$D(f_C\|f_B) = (2/3)\cdot\log\left(\frac{2/3}{1/2}\right) + (1/3)\cdot\log\left(\frac{1/3}{1/2}\right) \approx 0.025$$

Proposition 23 implies that the agent's beliefs converge to certainty of $\theta = A$.

7.2.2 Berk Nash Equilibrium

Standard equilibrium concepts in game theory assume that players best-respond to correct and common beliefs. Esponda and Pouzo (2016) proposes a new equilibrium concept (modifying Nash equilibrium) that allows players to be misspecified. As this definition can be applied also within a single-agent setting, and as the notation is substantially lighter in this case, we start by defining Berk Nash equilibrium with one agent.

Single Agent Settings

There is a finite set of payoff-relevant states Ω, a finite set of signal realizations $\mathbb{S}$, and a finite set of actions $\mathbb{A}$. The agent holds a prior p over $\Omega \times \mathbb{S}$. Additionally, there is a finite set of consequences $\mathbb{Y}$, which are determined by the agent's action and the state via a feedback function $f : \mathbb{A} \times \Omega \to \mathbb{Y}$. The agent's payoff function is $u : \mathbb{A} \times \mathbb{Y} \to \mathbb{R}$.

The timing is as follows. First the agent chooses a strategy $\sigma : \mathbb{S} \to \Delta(\mathbb{A})$ mapping the observed signal into a distribution over actions. Then, the state and signal (ω, s) are drawn according to p, and the action $\sigma(s)$ is implemented. Finally, the consequence y is determined given the action and state (a, ω), and the agent obtains payoff $u(a, y)$.

There is an *objective* mapping $Q : \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{Y})$ from actions and signals into distributions over consequences, where

$$Q(y \mid s, a) = \sum_{\omega : f(\omega, a) = y} p(\omega \mid s) \quad \forall (y, s, a).$$

This is the conditional distribution over consequences that a Bayesian agent with knowledge of f, the action a, and the signal realization s would expect.

The agent does not know Q (or f). His *subjective model* $\mathcal{Q} = \langle \Theta, (Q_\theta)_{\theta \in \Theta} \rangle$ is a parametrized family of mappings $Q_\theta : \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{Y})$.

Definition 7.2. *The agent is* correctly specified *if there exists $\theta \in \Theta$ such that $Q_\theta(\cdot \mid s, a) = Q(\cdot \mid s, a)$ for all $(s, a) \in \mathbb{S} \times \mathbb{A}$; otherwise the agent is* misspecified.

The following example is adapted from Esponda and Pouzo (2016):

Example 7.3. A monopolist chooses a price a, which together with a random shock $\omega \sim \mathcal{N}(0, 1)$ determines demand

$$y = f(a, \omega) = \phi(a) + \omega.$$

The monopolist's payoff is $u(a, y) = a \cdot y$. Under the objective mapping f , the conditional distribution $Q(\cdot \mid a)$ is normal with mean $\phi(a)$ and variance 1. The monopolist's subjective model is instead the family $Q_\theta(\cdot \mid a)$ of normal distributions indexed to $\theta = (\theta_0, \theta_1) \in \mathbb{R} \times \mathbb{R}$, where each $Q_\theta(\cdot \mid a)$ is normal with mean $\theta_0 + \theta_1 a$ and variance 1, corresponding to a perceived feedback function

$$f_\theta(a, \omega) = \theta_0 + \theta_1 a.$$

If ϕ is not in fact affine in a, then the monopolist is misspecified. (This example did not include a signal.)

For any agent strategy $\sigma : \mathbb{S} \to \Delta(\mathbb{A})$, define

$$q_\sigma(s, a) \equiv p_S(s) \sigma(a \mid s)$$

to be the distribution on $\mathbb{S} \times \mathbb{A}$ induced by the strategy σ and the agent's prior p. Further define

$$K(\sigma, \theta) = \sum_{(s,a) \in \mathbb{S} \times \mathbb{A}} \left(\mathbb{E}_{Q(Y \mid s, a)} \left[\ln \frac{Q(Y \mid s, a)}{Q_\theta(Y \mid s, a)} \right] \right) q_\sigma(a, s)$$

to be the expected Kullback-Leibler divergence between $Q_\theta(\cdot \mid s, a)$ and the objective distribution $Q(\cdot \mid s, a)$, weighted by $q_\sigma \in \Delta(\mathbb{S} \times \mathbb{A})$.

Given the agent's strategy σ, the set of closest parameters (in weighted KL divergence) is

$$\Theta^*(\sigma) = \arg\min_{\theta \in \Theta} K(\sigma, \theta)$$

Definition 7.3. *A strategy profile σ is a* Berk-Nash equilibrium *if there exists a $\mu \in \Delta(\Theta)$ such that*

(a) $\mu \in \Delta(\Theta^*(\sigma))$; i.e., μ has support on the set of KL-minimizers.

(b) σ is optimal given μ; namely, $\sigma(a \mid s) > 0$ implies that

$$a \in \arg\max_{a' \in \mathbb{A}} \mathbb{E}_{\overline{Q}_\mu(y|s,a')}[u(a', y)]$$

where $\overline{Q}_\mu(y \mid s, a) = \int_\Theta Q_\theta(y \mid s, a)\mu(\theta)d\theta$ is the conditional distribution over consequences that is induced by μ.

Example 7.4. A researcher's project is either good or bad, $\Omega = \{g, b\}$. The researcher observes a reaction to the project, which is either positive or negative, $\mathbb{S} = \{+, -\}$ where (ω, s) are jointly distributed according to:

	$s = +$	$s = -$
$\omega = g$	1/3	1/6
$\omega = b$	1/6	1/3

The researcher observes the signal $s \in \mathbb{S}$ and decides whether to exert high or low effort towards developing the project, $A = \{H, L\}$. The unknown true quality of the project, and the researcher's effort, jointly determine a journal outcome in $\mathbb{Y} = \{A, R\}$ (accept or reject) according to the following function

$$f(a, \omega) = \begin{cases} A & (a, \omega) = (H, g) \\ R & \text{otherwise} \end{cases}$$

That is, the project is accepted if it is good and also the researcher's effort is high, and it is rejected otherwise. The researcher's payoff is

$$u(a, y) = \begin{cases} 1 & (a, y) = (H, A) \\ -1 & (a, y) = (H, R) \\ 2 & (a, y) = (L, A) \\ 0 & (a, y) = (L, R) \end{cases}$$

The true distribution $Q(y \mid a, s)$ is described by $Q(A \mid +, L) = Q(A \mid -, L) = 0$ (since the paper will not be accepted if effort is low) and

$$Q(A \mid +, H) = p(\{\omega : f(H, \omega) = A\} \mid +) = p(g \mid +) = 2/3$$
$$Q(A \mid -, H) = p(\{\omega : f(H, \omega) = A\} \mid -) = p(g \mid -) = 1/3$$

since conditional on high effort, the probability of acceptance is equal to the probability that the paper is good. These conditional distributions are summarized as follows:

	A	R
$(+,H)$	$2/3$	$1/3$
$(-,H)$	$1/3$	$2/3$
$(+,L)$	0	1
$(-,L)$	0	1

Suppose the researcher's subjective model allows only for the parameters θ_1 and θ_2 which are indexed to the following conditional distributions:

	A	R		A	R
$(+,H)$	$3/4$	$1/4$	$(+,H)$	$2/3$	$1/3$
$(-,H)$	$1/2$	$1/2$	$(-,H)$	$1/3$	$2/3$
$(+,L)$	0	1	$(+,L)$	$1/10$	$9/10$
$(-,L)$	0	1	$(-,L)$	$1/10$	$9/10$

The distribution on the left, Q_{θ_1}, overestimates the value of hard work, and the distribution on the right, Q_{θ_2}, is overly optimistic about the probability of acceptance given low effort. Is the strategy profile $\sigma(+) = H$, $\sigma(-) = L$ (in which the research exerts high effort after a positive signal and low effort after a low signal) a Berk Nash equilibrium?

The distribution q_σ assigns probability 1/2 to $(+,H)$ and to $(-,L)$. So

$$K(\sigma,\theta) = \frac{1}{2}\left(\sum_{y\in\{A,R\}} Q(y \mid +,H)\cdot \ln\left(\frac{Q(y\mid +,H)}{Q_\theta(y\mid +,H)}\right)\right)$$
$$+\frac{1}{2}\left(\sum_{y\in\{A,R\}} Q(y \mid -,L)\cdot \ln\left(\frac{Q(y\mid -,L)}{Q_\theta(y\mid -,L)}\right)\right)$$

and thus

$$K(\sigma,\theta_1) = \frac{1}{2}\cdot\left(\frac{2}{3}\ln\left(\frac{2/3}{3/4}\right)+\frac{1}{3}\ln\left(\frac{1/3}{1/4}\right)\right) \approx 0.0038$$
$$K(\sigma,\theta_2) = \frac{1}{2}\cdot\ln\left(\frac{1}{9/10}\right) \approx 0.02$$

Hence θ_1 is the unique minimizer of KL divergence to the true distribution, i.e., $\Theta^*(\sigma) = \{\theta_1\}$.

Only $\mu = \delta_{\theta_1}$ (a point mass at θ_1) satisfies Part (a) of Definition 7.3, and the distribution $\overline{Q}_\mu$ in Part (b) of Definition 7.3 simplifies to Q_{θ_1}. To determine whether σ is a Berk Nash equilibrium, it remains to verify that σ satisfies the optimality condition in Part (b) of Definition 7.3.

Suppose the signal realization is $s = +$. Then the action H yields an expected payoff of

$$\mathbb{E}_{Q_{\theta_1}(y|+,H)}[u(H,y)] = 1 \cdot \frac{3}{4} - 1 \cdot \frac{1}{4} = \frac{1}{2}$$

while the action L yields an expected payoff of

$$\mathbb{E}_{Q_{\theta_1}(y|+,L)}[u(L,y)] = 0$$

so $a = H$ is indeed optimal.

Suppose the signal realization is $s = -$. Then the action H yields an expected payoff of

$$\mathbb{E}_{Q_{\theta_1}(y|-,H)}[u(H,y)] = 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0$$

while the action L yields an expected payoff of

$$\mathbb{E}_{Q_{\theta_1}(y|-,L)}[u(L,y)] = 0.$$

So $a = L$ is a best reply, and we conclude that σ is a Berk Nash equilibrium.

In sum, we have shown that the strategy σ is a best reply to a point mass on the unique parameter that minimizes KL divergence to the distribution over consequences induced by σ. In this sense the strategy σ is internally consistent with respect to the agent's misspecified model.

Exercise 7.1 (G). *Solve for all remaining pure-strategy Berk Nash equilibria in Example 7.4, or prove that there are none other.*

Simultaneous-Move Games

We turn now to the definition of Berk Nash equilibrium in simultaneous-move games. There is a set of players I, a set of payoff-relevant states Ω, a set of signal profiles $\mathbb{S} = \times_i \mathbb{S}_i$, and a probability distribution p over $\Omega \times \mathbb{S}$ whose marginals have full-support. There is a set of action profiles $\mathbb{A} = \times_i \mathbb{A}_i$, a set of *consequence* profiles $\mathbb{Y} = \times_i \mathbb{Y}_i$, and a profile of *feedback functions* $f = (f_i)_{i \in \mathcal{I}}$ where each $f_i : \mathbb{A} \times \Omega \to \mathbb{Y}_i$ maps outcomes in $\Omega \times \mathbb{A}$ into consequences for player i. Agents have payoff functions $u_i : \mathbb{A}_i \times \mathbb{Y}_i \to \mathbb{R}$.

The timing of the game is as follows: First, the state and signal (ω, s) are drawn according to p. Then each player i privately observes his own signal s_i and chooses an action a_i. The profile of consequences is determined via f as a function of the action profile and the state, and payoffs are realized.

For any player i, action $a_i \in \mathbb{A}_i$, and consequence $y_i \in \mathbb{Y}_i$, let

$$\Lambda^i(a_i, y_i) = \{(\omega, a_{-i}) : f_i(a_i, a_{-i}, \omega) = y_i\}$$

be the state and opponent action profiles that induce consequence y_i given player i's choice of a_i. The *objective distribution* over player i's consequences is $Q^i_\sigma : \mathbb{S}_i \times \mathbb{A}_i \to \Delta(\mathbb{Y}_i)$, where

$$Q^i_\sigma(y_i \mid s_i, a_i) = \sum_{(\omega, a_{-i}) \in \Lambda^i(a_i, y_i)} \sum_{s_{-i} \in S_{-i}} \left(\prod_{j \neq i} \sigma^j(a^j \mid s^j) \right) \cdot p_{\Omega \times S_{-i} | S_i}(\omega, s_{-i} \mid s_i)$$

for all $(s_i, a_i, y_i) \in \mathbb{S}_i \times \mathbb{A}_i \times \mathbb{Y}_i$. This is the conditional distribution over consequences that a Bayesian agent with knowledge of f, the strategy profile σ, and the signal realization s_i would expect.

The subjective model $\mathcal{Q} = \langle \Theta, (Q_\theta)_{\theta \in \Theta} \rangle$, with $\Theta = \prod_{i \in \mathcal{I}} \Theta^i$ and $Q_\theta = (Q^i_{\theta_i})_{i \in \mathcal{I}}$, describes the set of distributions over consequences that each player considers possible. Each player's parameter set Θ_i indexes distributions $Q^i_{\theta_i} : \mathbb{S}_i \times \mathbb{A}_i \to \Delta(\mathbb{Y}_i)$.

Definition 7.4. *A game is* correctly specified given σ *if for all players i, there exists $\theta_i \in \Theta_i$ such that $Q^i_{\theta_i}(\cdot \mid s_i, a_i) = Q^i_\sigma(\cdot \mid s_i, a_i)$ for all $(s_i, a_i) \in \mathbb{S}_i \times \mathbb{A}_i$; otherwise the game is* misspecified given σ. *A game is* correctly specified *if it is correctly specified for all σ; otherwise it is* misspecified.

For any strategy profile σ, define

$$q_{\sigma_i}(s_i, a_i) \equiv \sigma_i(a_i \mid s_i) p_{S_i}(s_i)$$

For any strategy profile σ, define

$$K_i(\sigma, \theta_i) = \sum_{(s_i, a_i) \in \mathbb{S}_i \times \mathbb{A}_i} \left(\mathbb{E}_{Q^i_\sigma(\cdot | s_i, a_i)} \left[\ln \frac{Q^i_\sigma(Y_i \mid s_i, a_i)}{Q^i_{\theta_i}(Y_i \mid s_i, a_i)} \right] \right) q_{\sigma_i}(s_i, a_i)$$

to be the expected Kullback-Leibler divergence between $Q_{\theta_i}(\cdot \mid s_i, a_i)$ and the objective distribution $Q^i_\sigma(\cdot \mid s_i, a_i)$, weighting (s_i, a_i) pairs according to $q_{\sigma_i}(s_i, a_i)$.

The set of closest parameters is

$$\Theta_i(\sigma) = \arg\min_{\theta_i \in \Theta_i} K_i(\sigma, \theta_i)$$

Definition 7.5. *A strategy profile σ is a* Berk-Nash equilibrium *if for all players i, there exists a $\mu_i \in \Delta(\Theta_i)$ such that*

(a) $\mu_i \in \Delta(\Theta_i(\sigma))$; i.e., μ has support on the set of KL-minimizers.

(b) σ_i is optimal given μ_i; namely, $\sigma_i(a_i \mid s_i) > 0$ implies that

$$a_i \in \arg\max_{\bar{a}_i \in \mathbb{A}_i} \mathbb{E}_{\overline{Q}^i_{\mu_i}(\cdot | s_i, \bar{a}_i)}[u_i(\bar{a}_i, Y_i)]$$

where $\overline{Q}^i_{\mu_i}(\cdot \mid s_i, \bar{a}_i) = \int_{\Theta_i} Q^i_{\theta_i}(\cdot \mid s_i, a_i) \mu_i(\theta_i) d\theta_i$ is the distribution over consequences of player i, conditional on $(s_i, a_i) \in \mathbb{S}_i \times \mathbb{A}_i$, induced by μ_i.

Remark 7.3. This definition is equivalent to Nash equilibrium when (a) is replaced with the condition that players have correct beliefs; i.e., $\overline{Q}^i_{\mu_i} = Q^i_\sigma$.

Proposition 24 (Esponda and Pouzo (2016))**.** *A Berk-Nash equilibrium exists.*

Building on Proposition 23, several authors have examined convergence of misspecified learning processes where—different from Berk (1966)'s setting—signals are endogenous to actions chosen by agents (Nyarko, 1991; Fudenberg, Romanyuk and Strack, 2017; Heidhues, Koszegi and Strack, 2021). The stable outcomes under many of these processes turn out to correspond to Berk Nash equilibria or a refinement of this set. Some recent works on this topic include Esponda and Pouzo (2016), Esponda, Pouzo and Yamamoto (2021), Bohren and Hauser (2021), Fudenberg, Lanzani and Strack (2020), Esponda, Pouzo and Yamamoto (2021) and Frick, Iijima and Ishii (2022).

7.3 Additional Exercises

Exercise 7.2 (G). *There are two states of the world, $\theta \in \{A, B\}$. A news source receives an infinite sequence of signals about this state of the world drawn iid according to the following signal structure*

	a	b
$\theta = A$	$3/4$	$1/4$
$\theta = B$	$1/4$	$3/4$

This news source is biased. When it observes the signal realization a, it reports a, but conditional on observing the signal realization b, it reports this b with probability $1 - \lambda$ and otherwise falsely reports a (where λ is constant across time). You are aware that the news source is biased and dogmatically believe that $\lambda = 1/2$.

Suppose the true state is $\theta = B$, and you observe the infinite sequence of news reports. Provide a condition (potentially empty) on the true value of λ such that your asymptotic belief is that the state is $\theta = A$. Interpret this result.

Chapter 8

Information Design

8.1 Bayesian Persuasion

8.1.1 Example

A judge and a prosecutor are involved in a court case. The unknown payoff-relevant state is whether the defendant in this case (who will not take an action) is *innocent* (I) or *guilty* (G). The judge and the prosecutor share a common prior that the defendant is guilty with probability 0.3.

The prosecutor cannot falsify or distort evidence, but can selectively choose what kind of information to present to the court (e.g., deciding who to subpoena or which forensic tests to conduct). Formally, the prosecutor chooses an information structure $\sigma : \{G, I\} \to \Delta(S)$ for some set of signal realizations S. The judge observes the outcome of the signal σ, updates his beliefs, and chooses whether to *acquit* or *convict* the defendant.

The judge and prosecutor's payoffs are determined by the judge's action and by the unknown state. The judge receives a payoff of 1 from convicting a guilty defendant or from acquitting an innocent defendant, and otherwise receives a payoff of zero. The prosecutor receives a payoff of 1 if the judge convicts the defendant and a payoff of 0 if the judge acquits the defendant, independent of the defendant's guilt. What information structure should the prosecutor choose, and what is the best expected payoff he can achieve?

Let's start with some benchmarks. One possibility is to send a completely uninformative signal. Since innocence is more likely than guilt under the judge's prior, the judge chooses to acquit given no information, yielding a payoff of zero for the prosecutor. Alternatively, the prosecutor can choose a perfectly informative signal that reveals the defendant's guilt. The judge convicts precisely when the defendant is guilty, yielding an expected payoff (under the prior) of 0.3 for the prosecutor.

Can the prosecutor do better? The perfectly revealing signal splits defendants into two bins—one labeled "convict" and one labelled "acquit" (Figure 8.1).

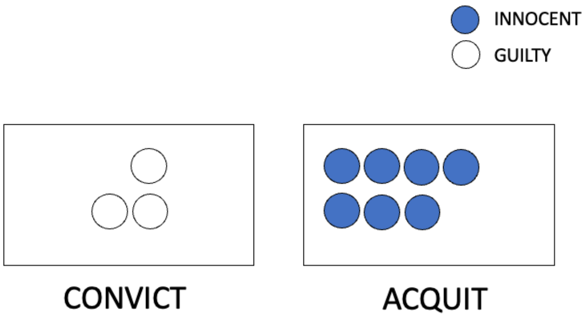

Figure 8.1: Depiction of the perfectly revealing signal, where each circle represents 1/10 of the population.

The judge's posterior for individuals labeled "convict" is that they are guilty with probability 1, so he optimally convicts any individual with this label. Likewise his posterior for individuals labeled "acquit" is that they are innocent with probability 1, so he acquits any individual with this label.

Now consider moving one unit of innocent individuals from the acquit bin to the convict bin (Figure 8.2).

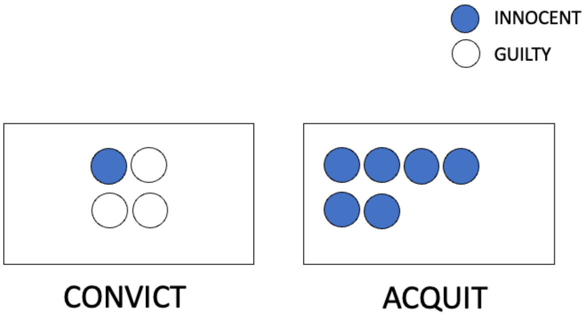

Figure 8.2: Deviation from the perfectly revealing signal.

Remark 8.1. Every "bin representation" as shown in Figures 8.1 and 8.2 corresponds to a unique signal. For each $\theta \in \Theta$ and $s \in \{\text{convict}, \text{acquit}\}$, let $P(\theta, s)$ be the mass of θ-type units in bin s (interpreting each circle as 1/10 of the population). Then P is a probability measure on $\Theta \times S$, and the corresponding signal $\sigma : \Theta \to \Delta(S)$ can be derived by Bayes' rule. As we see in the proof of Proposition 25, every signal also admits a bin representation.[1]

Following this modification on the perfectly revealing signal, the posterior probability of guilt in the acquit bin is unchanged. The posterior probability of

[1] In particular, every signal admits a "bin representation" that consists of two bins—a convict bin, and an acquit bin—where the judge optimally convicts all individuals in the convict bin and acquits all individuals in the acquit bin.

guilt for individuals labeled “convict” drops to 3/4—but crucially, the judge’s optimal action remains the same. Intuitively, by pooling innocent defendants with guilty defendants (but maintaining sufficiently guilty defendants that the judge still wants to convict), the prosecutor is able to induce the judge to wrongly convict a larger number of defendants.

Iterating this logic, we can continue moving units of innocent individuals into the convict bin, up until the judge is indifferent between convicting and acquitting (Figure 8.3).

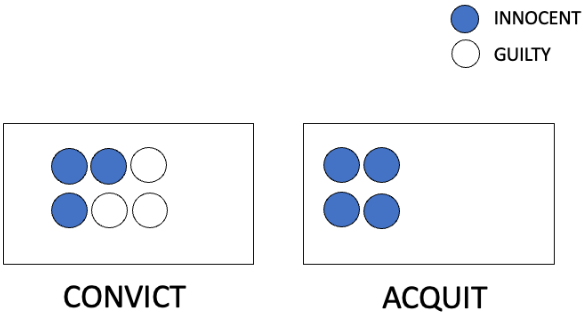

Figure 8.3: Depiction of the prosecutor-optimal signal structure.

These bins correspond to the following signal structure:

$$\begin{array}{ccc} & \textit{convict} & \textit{acquit} \\ G & 1 & 0 \\ I & 3/7 & 4/7 \end{array} \tag{8.1}$$

That this signal structure is optimal will follow from the results in the subsequent section. Strikingly, although the judge knows that only 30% of defendants are guilty, he ends up convicting 60% of them.

8.1.2 Model

There are two agents, a Sender and a Receiver. The unknown parameter θ takes values in the finite set Θ, and agents share a common prior $\mu_0 \in \Delta(\Theta)$. A signal is any mapping $\sigma : \Theta \to \Delta(S)$ from the set of states into distributions over a finite set of signal realizations S.

The Receiver chooses from a compact set of actions A. Both agents’ payoffs depend on the Receiver’s action and the unknown state. We’ll denote the Receiver’s utility function by $u_R : A \times \Theta \to \mathbb{R}$ and the Sender’s utility function by $u_S : A \times \Theta \to \mathbb{R}$, where both are assumed to be continuous.

The timeline is as follows: First, the Sender chooses a signal σ. The realization of this signal is then observed by the Receiver, who updates his beliefs and chooses an action $a \in A$. Finally payoffs are realized. The solution concept is

Sender-Preferred subgame perfect equilibrium; that is, the Receiver chooses an action to maximize his expected payoffs, breaking ties between optimal actions by maximizing Sender's payoffs.[2]

8.1.3 Solution and Geometric Representation

Consider any Sender-Preferred subgame perfect equilibrium, and let $\hat{a}(\mu)$ denote the Receiver's action given belief $\mu \in \Delta(\Theta)$ in this equilibrium. That is,

$$\hat{a}(\mu) \in \operatorname*{arg\,max}_{a \in A(\mu)} \mathbb{E}_\mu \left[u_S(a, \theta) \right] \tag{8.2}$$

where

$$A(\mu) = \operatorname*{arg\,max}_{a \in A} \mathbb{E}_\mu \left[u_R(a, \theta) \right]$$

is the set of actions that maximize the Receiver's expected payoff given belief μ. (If the RHS of (8.2) is non-empty, set $\hat{a}(\mu)$ to be any action in this set.) Let

$$\hat{v}(\mu) := \mathbb{E}_\mu \left[u_S(\hat{a}(\mu), \theta) \right] \tag{8.3}$$

be the Sender's expected payoff given belief μ and Receiver-action $\hat{a}(\mu)$. A signal's *value* is the Sender's (ex-ante) expected payoff given choice of that signal.

Proposition 25 (Kamenica and Gentzkow (2011)). *The following are equivalent:*

(i) There exists a (finite-valued) signal with value v^*.

(ii) There exists a (finite-valued) signal taking realizations in $S \subseteq A$ with value v^*.

(iii) There exists a Bayes-plausible distribution over posterior beliefs, $\tau \in \Delta(\Delta(\Theta))$, such that $\mathbb{E}_\tau \left[\hat{v}(\mu) \right] = v^*$.

Proof. The implication (ii) ⇒ (i) is immediate. The implication (ii) ⇒ (iii) follows from Fact 2.1 (every signal induces a Bayes-plausible distribution over posterior beliefs).

To show (i) ⇒ (ii), observe that for any signal $\sigma : \Theta \to \Delta(S)$ with value v^*, we can define a new signal $\widetilde{\sigma} : \Theta \to \Delta(A)$ that maps types into the recommended action under σ. That is,

$$\widetilde{\sigma}(a \mid \theta) = \sum_{s : \hat{a}(\mu_s) = a} \sigma(s \mid \theta)$$

for every $a \in A$ and $\theta \in \Theta$, where μ_s denotes the Receiver's posterior given signal realization s under σ. (The number of distinct action recommendations cannot exceed the size of S and so is finite.) Clearly the optimal action given recommendation of a remains the action a, so the distribution of optimal actions induced by $\widetilde{\sigma}$ and σ are the same.

[2] If there are multiple such actions, the Receiver chooses any action between them.

The direction (iii) ⇒ (i) is nearly immediate from Proposition 3 (every Bayes-plausible distribution over posterior beliefs can be induced by a signal), but we need to show that it is possible to construct a *finite-valued* signal for arbitrary τ (even ones with infinite support).[3]

We'll use the following result from convex analysis.

Proposition 26 (Caratheodory's Theorem). *Let $X \subseteq \mathbb{R}^n$ be a nonempty subset of finite-dimensional Euclidean space. Let $conv(X)$ denote the convex hull of X. Then every vector in $conv(X)$ can be represented as a convex combination of at most $n+1$ vectors from X.*

Fix any v^* and Bayes-plausible τ such that $\mathbb{E}_\tau[\hat{v}(\mu)] = v^*$. Define

$$C = \{(\mu, \hat{v}(\mu)) \mid \mu \in \Delta(\Theta)\}$$

to be the set of all beliefs and valuations of those beliefs, noting that $C \subseteq \mathbb{R}^n$ where $n \equiv |\Theta|$.[4] Moreover, by assumption that $v^* = \mathbb{E}_\tau[\hat{v}(\mu)]$ for some Bayes-plausible distribution τ over posterior beliefs, the vector (μ_0, v^*) belongs to the convex hull of C.

Then by Caratheodory's Theorem, there exists a sequence of beliefs $(\mu_i)_{i=1}^{n+1}$ and a sequence of nonnegative weights $(\alpha_i)_{i=1}^{n+1}$ summing to 1, such that

$$(\mu_0, v^*) = \sum_{i=1}^{n+1} \alpha_i \cdot (\mu_i, \hat{v}(\mu_i))$$

Let τ^* be the distribution over posterior beliefs that assigns probability α_i to each belief μ_i, $1 \leq i \leq n+1$. Then

$$\mathbb{E}_{\tau^*}[\hat{v}(\mu)] = \sum_{i=1}^{n+1} \alpha_i \cdot \hat{v}(\mu_i) = v^*$$

as desired. Follow the construction in Section 2.2.2 (setting the set of signal realizations S to be the posterior beliefs in the support of τ^*) to complete the proof. ■

Proposition 25 tells us that we can determine when the Sender benefits from persuasion by studying how $\mathbb{E}_\tau\left[\hat{v}(\mu)\right]$ varies over the set of Bayes-plausible distributions.

Corollary 8.1. *The Sender benefits from persuasion if and only if there exists a Bayes-plausible distribution τ such that $\mathbb{E}_\tau\left[\hat{v}(\mu)\right] > \hat{v}(\mu_0)$.*

Corollary 8.2. *The value of an optimal signal is*

$$\max_{\tau \in \Delta(\Theta)} \mathbb{E}_\tau\left[\hat{v}(\mu)\right] \quad s.t. \int \mu d\tau(\mu) = \mu_0$$

[3] The construction in Section 2.2.2 chooses S to be the set of all beliefs in the support of τ, which need not be finite.

[4] The simplex $\Delta(\Theta)$ is a subset of $\mathbb{R}^{n-1}$ and the valuation belongs to $\mathbb{R}$, hence $C \subseteq \mathbb{R}^n$.

The value of information for the Sender at any prior μ can be represented geometrically using the upper concave envelope of $\hat{v}$.

Definition 8.1. *Define*

$$V(\mu) \equiv \sup\{z \mid (\mu, z) \in Conv(\hat{v})\} \quad \forall \mu \in \Delta(\Theta)$$

where $Conv(\hat{v})$ denotes the convex hull of the graph $\hat{v}$. That is, V is the smallest concave function that is everywhere weakly greater than $\hat{v}$.

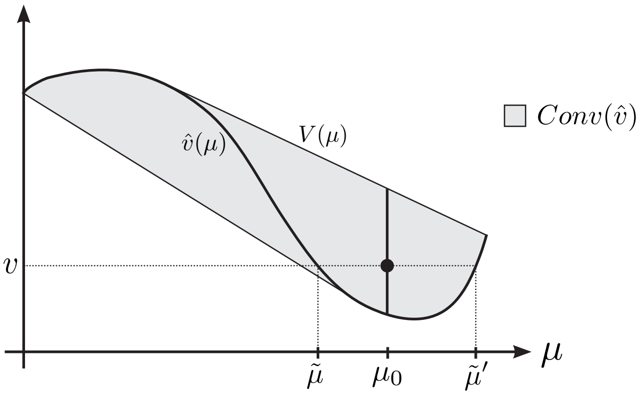

Figure 8.4: Illustration of Definition 8.1.

By Proposition 25, the set $\{z \mid (\mu_0, z) \in Conv(\hat{v})\}$ is precisely those expected payoffs that the Sender can achieve when the prior μ_0. For example, in Figure 8.1, the value v is achievable from the prior μ_0 via a signal that splits the prior into two posterior $\tilde{\mu}$ and $\tilde{\mu}'$ (setting the weights so that the expected posterior equals the prior). So $V(\mu_0) = \sup\{z \mid (\mu_0, z) \in Conv(\hat{v})\}$ is the largest payoff Sender can achieve when the prior is μ_0, and the Sender strictly benefits from persuasion if and only if $V(\mu_0) > \hat{v}(\mu_0)$.

The following corollary is immediate from the previous analysis.

Corollary 8.3. *If $\hat{v}$ is concave, then the Sender does not benefit from persuasion for any prior. If $\hat{v}$ is strictly convex, the Sender benefits from persuasion for every prior.*

8.1.4 Back to the Example

Returning to the setting of Section 8.1.1, observe that in any Sender-preferred subgame equilibrium, the judge's action given probability of guilt μ is

$$\hat{a}(\mu) = \begin{cases} \textit{convict} & \text{if } \mu \geq 0.5 \\ \textit{acquit} & \text{if } \mu < 0.5 \end{cases}$$

where the tie at $\mu = 0.5$ is broken in favor of the prosecutor. So the prosecutor's expected payoff is

$$\hat{v}(\mu) = \begin{cases} 1 & \text{if } \mu \geq 0.5 \\ 0 & \text{if } \mu < 0.5 \end{cases}$$

as depicted in Panel (a) of Figure 8.5.

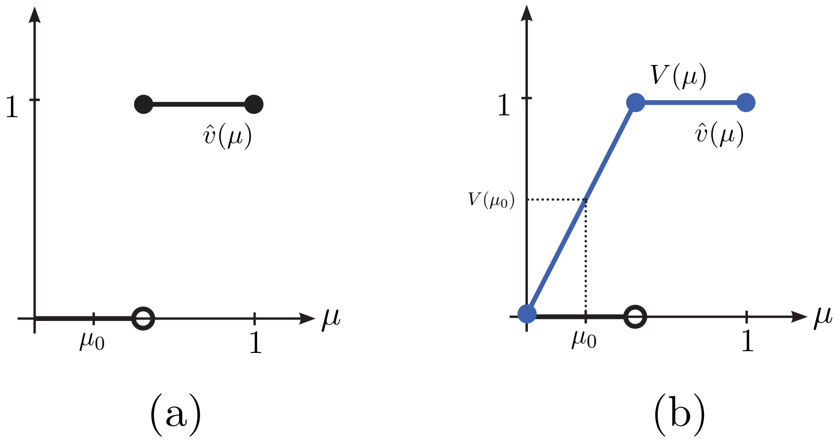

Figure 8.5: Depiction of $\hat{v}(\mu)$ in the prosecutor-judge example.

The upper concave envelope of $\hat{v}$ is

$$V(\mu) = \begin{cases} 1 & \text{if } \mu \geq 0.5 \\ 2\mu & \text{if } \mu < 0.5 \end{cases}$$

as depicted in Panel (b) of Figure 8.5. At the prior belief of $\mu_0 = 0.3$, we have $V(0.3) = 0.6$, confirming that the signal structure in (8.1) delivers the best possible expected payoff for the prosecutor. We see moreover that the prosecutor benefits from persuasion whenever $\mu_0 < 0.5$ (i.e., whenever the judge would optimally acquit under the prior), but cannot improve his expected payoff through choice of any signal structure when $\mu_0 \geq 0.5$.

8.2 Additional Exercises

Exercise 8.1 (U). *A student (Sender)'s quality is $\theta \in \{L, H\}$. The employer chooses an action from $A = \{l, m, h\}$ where l is a low-responsibility position, m is a medium-responsibility position, and h is a high-responsibility position. The employer's payoffs are:*

$$u_E(a, \theta) = \begin{cases} 1 & \textit{if } (a, \theta) = (H, h) \\ 0 & \textit{if } (a, \theta) \in \{(H, m), (H, l), (L, l)\} \\ -1 & \textit{if } (a, \theta) \in \{(L, m), (L, h)\} \end{cases}$$

The student's (state-independent) payoff function u_S takes value 1 if $a = h$, 0 if $a = m$, and -1 if $a = l$.

(a) Suppose the employer's beliefs are described as $(p, 1-p)$, where p is the probability of $\theta = L$. Let

$$\hat{a}(p) = \arg\max_{a \in \{l,m,h\}} \mathbb{E}_{(p,1-p)}[u_E(a,\theta)].$$

(This is the same as in (8.2), except we simplify notation by writing $\hat{a}(p)$ instead of $\hat{a}(p, 1-p)$.) Solve for $\hat{a}(p)$ on the domain $p \in [0,1]$, assuming that the employer breaks ties in favor of the action that maximizes the student's payoffs.

(b) Suppose the student's beliefs are described as $(p, 1-p)$, where p is the probability of $\theta = L$, and the student knows that the employer chooses action $\hat{a}(p)$. Let

$$\hat{v}(p) = \mathbb{E}_{(p,1-p)}[u_S(\hat{a}(p), \theta)]$$

denote the student's expected payoff at this belief. Solve for $\hat{v}(p)$ on the domain $p \in [0,1]$ and plot it.

(e) Let $V(p)$ be the smallest concave function that is everywhere above $\hat{v}(p)$. Reproduce your plot from part (d) with $V(p)$ and $\hat{v}(p)$ depicted in the same figure. Clearly label $V(p)$ and $\hat{v}(p)$.

(f) Identify all $p \in [0,1]$ such that $V(p) > \hat{v}(p)$. These are the prior beliefs at which the student can strictly benefit from design of the signal structure.

Exercise 8.2 (G). *Fix an arbitrary finite set of states Θ and finite set of actions A. Suppose that the Sender and Receiver's payoff functions satisfy*

$$u_S(a,\theta) = -u_R(a,\theta)$$

for every $a \in A$ and $\theta \in \Theta$. Prove that $V(\mu) = \hat{v}(\mu)$ for every belief μ, where $\hat{v}$ is as defined in (8.3) and V is as given in Definition 8.1. Interpret this result.

Bibliography

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022. “Too Much Data: Prices and Inefficiencies in Data Markets.” *AEJ: Microeconomics.*

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz. 2015. “Fragility of Asymptotic Agreement under Bayesian Learning.” *Theoretical Economics*, 11(1): 187–225.

Arrow, K. J., D. Blackwell, and M. A. Girshick. 1949. “Bayes and Minimax Solutions of Sequential Decision Problems.” *Econometrica*, 17(3/4): 213–244.

Aumann, Robert J. 1976. “Agreeing to Disagree.” *The Annals of Statistics*, 4(6): 1236–1239.

Bagnoli, Mark, and Ted Bergstrom. 2005. “Log-concave probability and its applications.” *Economic Theory*, 26(none): 445 – 469.

Balakrishna, N., and Chin Diew Lai. 2009. “Concepts of Stochastic Dependence.” *Continuous Bivariate Distributions: Second Edition*, 105–140. New York, NY:Springer New York.

Berk, Robert H. 1966. “Limiting Behavior of Posterior Distributions when the Model is Incorrect.” *The Annals of Mathematical Statistics*, 37(1): 51 – 58.

Blackwell, David. 1951. “Comparison of Experiments.” *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, 930–102.

Blackwell, David, and Lester Dubins. 1962. “Merging of Opinions with Increasing Information.” *The Annals of Mathematical Statistics.*

Bloedel, Alexander, and Weijie Zhong. 2021. “The Cost of Optimally-Acquired Information.” Working Paper.

Bohren, Aislinn, and Daniel Hauser. 2021. “Learning with Model Misspecification: Characterization and Robustness.” *Econometrica*, 89: 3025–3077.

Bregman, Lev. 1967. “The Relaxation Method of Finding Common Points of Convex Sets and Its Application to the Solution of Problems in Convex Programming.” *USSR Computational Mathematics and Mathematical Physics*, 7(3).

Brooks, Benjamin, Alexander Frankel, and Emir Kamenica. 2022*a*. "Comparisons of Signals." Working Paper.

Brooks, Benjamin, Alexander Frankel, and Emir Kamenica. 2022*b*. "Information Hierarchies." *Econometrica*.

Caplin, Andrew, and Mark Dean. 2013. "Behavioral Implications of Rational Inattention with Shannon Entropy." Working Paper.

Caplin, Andrew, Mark Dean, and John Leahy. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review*, 105(7).

Caplin, Andrew, Mark Dean, and John Leahy. 2022. "Rationalizing Inattentive Behavior: Characterizing and Generalizing Shannon entropy." *Journal of Political Economy*, 130(6).

Chambers, Christopher P., and Paul J. Healy. 2011. "Reversals of Signal-Posterior Monotonicity For Any Bounded Prior." *Mathematical Social Sciences*, 61(3): 178–180.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data*, 5(2): 153–163.

Cripps, Martin, Jeffrey Ely, George Mailath, and Larry Samuelson. 2008. "Common Learning." *Econometrica*, 76(4): 909–933.

de Castro, Luciano. 2009. "Affiliation and Dependence in Economic Models." Working Paper.

Denti, Tomasso. 2022. "Posterior Separable Cost of Information." *American Economic Review*, 112(10).

de Oliveira, Henrique. 2019. "Blackwell's Informativeness Theorem Using Diagrams." *Games and Economic Behavior*, 109: 126–131.

Doob, Joseph. 1949. "Application of the theory of martingales." *Le Calcul des Probabilités et ses Applications*, 23–27. Springer New York.

Esary, J. D., F. Proschan, and D. W. Walkup. 1967. "Association of Random Variables, with Applications." *The Annals of Mathematical Statistics*, 38(5): 1466 – 1474.

Esponda, Igancio, Demian Pouzo, and Yuichi Yamamoto. 2021. "Limit Points of Endogenous Misspecified Learning." *Journal of Economic Theory*, 195.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica*, 84(3): 1093–1130.

Frankel, Alexander, and Emir Kamenica. 2019. "Quantifying Information and Uncertainty." *American Economic Review*, 109(10): 3650–3680.

Frick, Mira, Ryota Iijima, and Yuhta Ishii. 2022. "Belief Convergence under Misspecified Learning: A Martingale Approach." Forthcoming.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack. 2020. "Limit Points of Endogenous Misspecified Learning." Working Paper.

Fudenberg, Drew, Gleb Romanyuk, and Philipp Strack. 2017. "Active Learning with a Misspecified Prior." *Theoretical Economics*, 12: 1155–1189.

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, Accuracy, and the Optimal Timing of Choices." *American Economic Review*, 108(12): 3651–84.

Geanakoplos, John, and Heraklis Polemarchakis. 1982. "We Can't Disagree Forever." *Journal of Economic Theory*, 28(1): 192–200.

Hébert, Benjamin, and Jennifer La'O. 2022. "Information Acquisition, Efficiency, and Non-Fundamental Volatility." Working Paper.

Hebert, Benjamin, and Michael Woodford. 2021*a*. "Neighborhood-Based Information Costs." *American Economic Review*, 111(10).

Hebert, Benjamin, and Michael Woodford. 2021*b*. "Rational Inattention When Decisions Take Time." Working Paper.

Heidhues, Paul, Botond Koszegi, and Philipp Strack. 2021. "Convergence in models of misspecified learning." *Theoretical Economics*, 16(1): 73–99.

Heinsalu, Sander. 2020. "Reversals of signal-posterior monotonicity imply a bias of screening." *Journal of Economic Theory*, 188.

Holmstrom, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *The Review of Economic Studies*, 66(1): 169–182.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasian." *American Economic Review*, 101(6): 2590–2615.

Kartik, Navin, Frances Lee, and Wing Suen. 2021. "Information Validates the Prior: A Theorem on Bayesian Updating and Applications." *American Economic Review: Insights*, 3(2).

Khinchin, A.I. 1957. *Mathematical Foundations of Information Theory.* AMLBook.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ITCS*.

Lagziel, David, and Ehud Lehrer. 2019. "A Bias of Screening." *AER: Insights*, 1(3): 343–356.

Lehmann, E. L. 1966. "Some Concepts of Dependence." *The Annals of Mathematical Statistics*, 37(5): 1137 – 1153.

Leshno, Moshe, and Yishay Spector. 1992. “An Elementary Proof of Blackwell’s Theorem.” *Mathematical Social Sciences.*

Liang, Annie, Xiaosheng Mu, and Vasilis Syrgkanis. 2022. “Dynamically Aggregating Diverse Information.” *Econometrica*, 90(1): 47–80.

Matějka, Filip, and Alisdair McKay. 2015. “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Login Model.” *American Economic Review*, 105(1).

Meyer, Margaret A. 1991. “Learning from Coarse Information: Biased Contests and Career Profiles.” *The Review of Economic Studies*, 58(1): 15–41.

Milgrom, Paul. 1981. “Good News and Bad News: Representation Theorems and Applications.” *The Bell Journal of Economics*, 12(2): 380–391.

Miller, Jeffrey W. 2018. “A Detailed Treatment of Doob’s Theorem.”

Monderer, Dov, and Dov Samet. 1989. “Approximating Common Knowledge with Common Beliefs.” *Games and Economic Behavior*, 1: 170–190.

Morris, Stephen, and Hyun Song Shin. 2002. “Social Value of Public Information.” *American Economic Review*, 92(5): 1521–1534.

Morris, Stephen, and Philipp Strack. 2019. “The Wald Problem and the Equivalence of Sequential Sampling and Static Information Costs.” Working Paper.

Nyarko, Yaw. 1991. “Learning in mis-specified models and the possibility of cycles.” *Journal of Economic Theory*, 55(2): 416–427.

Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2020. “The Cost of Information.” Working Paper.

Rubinstein, Ariel. 1989. “The Electronic Mail Game: Strategic Behavior Under Almost Common Knowledge.” *American Economic Review*, 79(3): 385–391.

Saumard, Adrien, and Jon A. Wellner. 2014. “Log-concavity and strong log-concavity: A review.” *Statistics Surveys*, 8(none): 45 – 114.

Shannon, Claude Elwood. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal*, 27: 379–423.

Shmaya, Eran, and Leeat Yariv. 2016. “Experiments on Decisions under Uncertainty: A Theoretical Framework.” *American Economic Review*, 106(7): 1775–1801.

Sims, Christopher. 2003. “Implications of Rational Inattention.” *Journal of Monetary Economics*, 50(3).

Wald, Abraham. 1945. “Sequential Tests of Statistical Hypotheses.” *The Annals of Mathematical Statistics*, 16(2): 117–[186.