UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Information in Dynamic Games and Mechanism Design

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Daehyun Kim

2019

ABSTRACT OF THE DISSERTATION

Essays on Information in Dynamic Games and Mechanism Design

by

Daehyun Kim
Doctor of Philosophy in Economics
University of California, Los Angeles, 2019
Professor Ichiro Obara, Chair

This dissertation studies how asymmetric information between economic agents interacts with their incentive in dynamic games and mechanism design. Chapter 1 and Chapter 2 study this in mechanism design, especially focusing on robustness of mechanisms when a mechanism designer's knowledge on agents' belief and higher order beliefs is not perfect. In Chapter 1 we introduce a novel robustness notion into mechanism design, which we term confident implementation; and characterize confidently implementable social choice correspondences. In Chapter 2, we introduce another robust notion, $\mathbf{p}$-dominant implementation where $\mathbf{p} \in [0,1]^N$ and $N \in \mathbb{N}$ is the number of agents, and fully characterize $\mathbf{p}$-dominant implementable allocations in the quasilinear environment. Chapter 1 and Chapter 2 are related in the following way: for some range of $\mathbf{p}$, a $\mathbf{p}$-dominant implementable social choice correspondence is confidently implementable.

In Chapter 3, we study information disclosure problem to manage reputation. To study this, we consider a repeated game in which there are a long-run player and a stream of short-run players; and the long-run player has private information about her type, which is either *commitment* or *normal*. We assume that the shot-run player only can observe the past $K \in \mathbb{N}$ periods of information disclosed by the long-run player. In this environment, we characterize the information disclosure behavior of the long-run player and also equilibrium dynamics whose shape critically depends on the prior.

The dissertation of Daehyun Kim is approved.

Sushil Bikhchandani

Joseph M. Ostroy

Moritz Meyer-ter-Vehn

Ichiro Obara, Committee Chair

University of California, Los Angeles

2019

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I am deeply indebted to my advisor, Ichiro Obara. Having him as my advisor is a great honor and the best thing I did at UCLA. He is not only a brilliant mind but also a warmhearted person. From the countless meetings we had since the last quarter of my second year, I learned many things from him, varying from knowledge on various topics and research skill to his attitude toward students and the profession.

I am also extremely grateful to Moritz Meyer-ter-Vehn. He has always tried to find positive aspects of my researches and me. Such a feeling has been truly encouraging throughout my Ph.D. period. Although it was challenging, to prepare myself to convince him about my research was extremely helpful (and enjoyable). I have been so fortunate to have him in my committee.

I also would like to express my deepest appreciation to my other committee members. It is a great honor to have had conversations with Joseph Ostroy, who has been an exemplar of an economist to me, glancing at his original view to understand the world. Also, Sushil Bikhchandani has generously shared his valuable time for serving as a member of the committee and for helpful conversations. I wish I would have met them more often. I also would like to thank the other UCLA theory group: Simon Board, Tomasz Sadzik, Jay Lu, John Riley, Marek Pycia for their insightful advices during my presentations and conversations.

Yingju Ma has been a very close friend since we met first. I will not forget many things we did together during last few years. I want to also thank Rustin Partow for share his insight and effort for our ongoing project. I learned many things from him. The life in the school became much more enjoyable after I began to share the office with Diana Van Patten and Renato Giroldo. I also want to thank my peer and friend Xinyu Fan, Zhouran Lu, Jonathan Gu, Jinwook Hur, Yun Feng, Keyoung Lee, YongKi Hong, and Jeonghwan Kim.

Last, but not least, I would like to thank my parents, brother, sister-in-law and my niece. Without their unconditional support, I would not be able to finish this challenge. I will always remember Byul (2002–2017) who had stayed with my family for a long time.

VITA

2011	Bachelor of Business Administration (with Highest Honors) Yonsei University, Seoul, South Korea
2013	M.A. in Economics Yonsei University, Seoul, South Korea
2014	M.A. in Economics University of California, Los Angeles
2015–2019	C.Phil. in Economics University of California, Los Angeles
2014–2019	Teaching Assistant, Associate, and Fellow Department of Economics, University of California, Los Angeles

PUBLICATIONS

Kim, D. (2019). Comparison of information structures in stochastic games with imperfect public monitoring. *International Journal of Game Theory 48*(1), 267–285.

CHAPTER 1

Confident Implementation

1.1 Introduction

In mechanism design theory, a type space (Harsanyi, 1967, 1968a,b) is used to model agents belief and higher order beliefs; and it is implicitly or explicitly assumed that the type space is common knowledge (Lewis, 1969; Aumann, 1976) between the agents and the designer of a mechanism. This means that the designer is *certain* about each agent's possible private information about payoff-relevant parameters (called *payoff type*); and also certain about agents' possible beliefs over other agents' payoff types, *ad infinitum.* Such a situation is regarded as an idealization or approximation of reality especially when the type space is "small."[1] On the other hand, in the relatively recent literature on robust mechanism design, the "global" approach is mostly employed where a mechanism is required to be robust enough that in *any* type space involved incentive compatibilities are satisfied.[2] An implication of this approach is that the designer does not need to have any information about agents.

However, in reality, the designer often has some information from some investigation about agents with substantial accuracy; at the same time, the designer is hardly certain about this information, worrying about some unexpected situations in which such information turns out to be not true; therefore the designer may want to design a mechanism to be robust with respect to such concern. For instance, the designer is "quite sure" about agents' possible

[1] See for example, Monderer and Samet (1989). On the other hand, assuming the common knowledge is without loss in a particularly "large" type space called the universal type space, which shall be discussed later in detail.

[2] There are other researches to take a certain notion of localness. See the literature review which follows shortly.

first order beliefs, while admitting that agents might have *some* unexpected first order belief with a small probability.

We introduce a novel framework to study mechanism design problems to take into account this reality. Imagine the situation in which the designer has information about each agent's possible first order beliefs. One may consider the approach to model it by first introducing a particular type space whose induced first order beliefs coincide with the designer's information; then assuming that this type space is likely to happen. However, note that this approach also means that the probability of the event in which such information of the first order beliefs is *common knowledge* among agents is also likely, which is not part of the designer's original information. We provide an approach to circumvent this problem: we do not impose a *particular* type space to describe the designer's information; rather we model it as an *event* in the *universal type space* (Harsanyi, 1967; Mertens and Zamir, 1985; Brandenburger and Dekel, 1993) which consists of all possible coherent belief hierarchies and assuming its common knowledge among agents and the designer is without loss. Note that an event in the universal type space is generally not belief closed.

Given this, how do we model the designer who is "quite sure" about her information? In this paper, we take the following approach: whichever a type space turns out to be true, in the type space the event is sufficiently likely to happen. In other words, the designer considers every type space in which the event is sufficiently likely to happen.

Our novel local robustness notion which we term *confident implementation* requires that a mechanism "approximately" implements a social choice correspondence in any countable common prior type spaces in which the event is sufficiently likely to happen, but not necessarily probability 1. Here approximation means that in any such type spaces, there is an equilibrium such that it achieves an element of the social choice correspondence with probability arbitrarily close to 1. One might think if an event is sufficiently likely, then the outcome should be similar to when the event is certain. This is not necessarily the case as it is pointed out by Rubinstein (1989).

After establishing our framework and robustness notion, we characterize social choice correspondences that are confidently implementable with respect to some information given to the designer, which is modeled a subset E^* in the universal type space as we mentioned above. Especially, we focus on the situation where the designer has information about agents' payoff type and n-th order belief for some finite $n \in \mathbb{N}$. Under a condition, which we term *distinguishability*, we characterize a subset of social choice correspondence that are confidently implementable with respect to this information; namely **p**-dominant implementable social choice correspondences where $\mathbf{p} \in [0,1]^N$ and $\sum p_i < 1$. The notion of **p**-dominant equilibrium (Morris et al., 1995; Kajii and Morris, 1997) is extended to incomplete information games for this purpose. We show that if a social choice correspondence is **p**-dominant implementable, where $\sum_i p_i < 1$, in the *maximal* consistent belief closed subset, which includes all the consistent belief closed subsets in E^*, then it is confidently implementable with respect to E^*.

We believe that our local notion of robustness and the sufficient characterization results are important because of the following two existing results: First, Bergemann and Morris (2005) shows that a social choice *function* is globally robust implementable if and only if it is ex-post implementable. Second, Jehiel et al. (2006) shows that a social choice function is ex-post implementable if and only if it is constant under a mild assumption.[3] Together they imply that globally robust implementable social choice functions are extremely limited. Our local notion and the result may open some possibility of robustly implementing social choice functions as long as the designer's information is sufficiently accurate.

It is also useful to compare our robustness notion to the existing ones more carefully. Bergemann and Morris (2005) studies (partial) implementation problem in the situation where a mechanism designer is assumed to know nothing about agents' payoff environment. Thus, their approach is "global" in the sense that the designer wants to make a mechanism that works for any agents' beliefs and higher order beliefs; in other words, agents' incentive compatibilities hold in the universal type space.[4] On the other hand, Oury and Tercieux

[3] To be more precise, if agents' signals are at least two-dimensional.

[4] Bergemann and Morris (2009a) studies full implementation problem using direct mechanisms with the solution concept of rationalizability. They study general mechanisms in Bergemann and Morris (2011). They

(2012) study a local concept of robustness in the following sense. They study conditions for social choice functions to be approximately implemented when agents' belief hierarchy is sufficiently close to the benchmark hierarchy in terms of the product topology on the universal type space. In this sense, their approach is interim, while ours is ex-ante.[5] Perhaps, the most closest robustness to ours is that of Artemov et al. (2013). They suggest a local robustness notion which captures situations in which a certain set of first order beliefs are assumed to be common knowledge between agents and the designer (see also Jehiel et al. (2012); Lopomo et al. (2009); Ollár and Penta (2017)). Although we also study the designer has some information about agents' belief hierarchies; but we do *not* assume any common knowledge of this information. Namely, in the case of information regarding first order beliefs, *any* first order beliefs might happen with a small chance. In this sense, out local notion is probabilistic. Lastly, Meyer-ter-Vehn and Morris (2011) considers a situation where an assumed utility function (in addition to its parameters) may not be accurate or not common knowledge.

This paper is also related to the literature on robustness of equilibrium. Especially, in terms of concept of closeness, ours is similar to Kajii and Morris (1997) in which they study robustness of Nash equilibrium in complete information game by informationally perturbing complete information games.[6] As a sufficient condition of robustness they suggest a Nash equilibrium to be **p**-dominant with a certain range of **p**.[7,8] We extend this solution concept to games with incomplete information; and also introduce a partial implementation notion using the extended solution concept.

Some papers study mechanisms that capture different types of robustness. For instance,

study a similar question with virtual implementation in Bergemann and Morris (2009b)

[5] See also Weinstein and Yildiz (2007).

[6] See also Kajii and Morris (1998). Given a game-form and state space, they study conditions for two distributions over states by which for any utility function and any equilibrium with one distribution, there exists an approximate equilibrium which gives approximately the same ex-ante payoff with the other distribution.

[7] Morris et al. (1995) first introduce this equilibrium concept to study conditions for "infection argument" to work.

[8] For non-common prior perturbation see, Oyama and Tercieux (2010) where they show that a game has a "robust" equilibrium if and only if this game is dominance solvable.

Eliaz (2002) considers situations where some agents are irrational which he call *faulty*. His k-tolerant implementation is immune to the situations where at most k agents are faulty. As we will discuss, $\mathbf{p}$-dominant implementation can be interpreted as a robust implementation notion to capture such a situation, while with a different modeling of the possibility of faulty agents.

The remaining of the paper is organized as follows: In Section 1.2, we provide an example by which we (informally) explain our notion of robustness and framework; then we show that there is a social choice correspondence that is not ex-post implementable but robustly implementable in our sense. In Section 1.3, we describe the environment of our mechanism design problem. Then in Section 1.4, we formally introduce our robustness notion called confident implementation and framework to study this. We also introduce $\mathbf{p}$-dominant implementation, which shall play an important role in our characterization of confident implementability. Throughout Section 1.5, our main results are illustrated with the special cases involving the first order belief and second order belief information. In Section 1.6, we provide our main result which characterizes confident implementability when the designer has information about n-th order belief of agents for any finite $n \in \mathbb{N}$. In Section 1.7, we discuss an additional sufficient condition for confident implementation. Lastly in Section 1.8, we conclude the paper with discussion of future directions.

1.2 An Example

To make our discussion more concrete, consider the following example:

Example 1.1. There is a public good which the designer is considering to build. Let $X = \{0, 1\}$ where $x = 1$ and $x = 0$ represents building the public good and not building it, respectively; let $\mathcal{X} = \Delta(X) = [0, 1]$, the probability of $x = 1$. There are 2 agents who have interdependent valuation: for each i let $v_i(\theta) := \theta_i + \theta_j$, $j \neq i$, be the value of the public good to agent i if $x = 1$; and let $\Theta_i = \{\theta_h, \theta_l\}$ where $\theta_h = 1, \theta_l = -2$.[9]

[9] Note that there is no money to transfer utility in this example.

F	θ_h	θ_l
θ_h	1	1/2
θ_l	1/2	0

Figure 1.1: Social choice function for Example 1.1

Suppose that the designer wants to implement the social choice correspondence (in fact, it is a function) in 1.1.

The designer acquires information indicating that each agent i's possible pairs of payoff profile and the first order belief is *likely* to be in the following set:

$$\Delta_i^1 \equiv \{(\theta_h, \lambda_i^{\theta_h}), (\theta_l, \lambda_i^{\theta_l})\}$$

where $\lambda_i^{\theta_i} \in \Delta(\{\theta_h, \theta_l\})$ for each i and θ_i. Let $\Delta^1 \equiv \Delta_1^1 \times \Delta_2^1$. Note that the designer's information does not involve any higher order beliefs beyond the first order. Note that the designer's information has the following property: for each θ_i there is only one first order belief based on her information. In fact, this is implicitly assumed in the standard type space (*payoff type spaces*), which is mostly employed in the mechanism design literature and also applied models. We also use such particular first order beliefs in this example to emphasize its relation to the standard type space, although our results do not involve such a restriction. We further assume in this example, for simplicity, that $\lambda_i^{\theta_i} = (\lambda, 1 - \lambda)$ for each i and θ_i where $\lambda \in [0, 1]$ is the probability that each agent, independently of his type, believes the other agent is θ_h.

We should emphasize that the designer cannot be *certain* of this information and this is the most important distinction from the conventional approach. Namely, with a small probability, in the true world (true type space), agents may have any other payoff types or other first order beliefs (or both). In contrast, when the designer is *certain* of her information (as it is in the conventional approach), the unique possible type space is the one described in 1.2. We denote this type space by $((\Theta)_{i \in \mathcal{I}}, \hat{\lambda})$.[10]

[10] Our result does not involve the uniqueness of type space when the designer is certain.

$\hat{\lambda}$	θ_h	θ_l
θ_h	λ^2	$\lambda(1-\lambda)$
θ_l	$\lambda(1-\lambda)$	$(1-\lambda)^2$

Figure 1.2: The type space when the designer is *certain* in Example 1.1

β	θ_h, s_2'	θ_h, s_2''	θ_l, s_2'	θ_l, s_2''
θ_h, s_1'	$\frac{1}{4}(1-\epsilon)$	$\frac{1}{12}\epsilon$	$\frac{1}{4}(1-\epsilon)$	$\frac{1}{12}\epsilon$
θ_h, s_1''	$\frac{1}{12}\epsilon$	0	$\frac{1}{12}\epsilon$	$\frac{1}{6}\epsilon$
θ_l, s_1'	$\frac{1}{4}(1-\epsilon)$	$\frac{1}{12}\epsilon$	$\frac{1}{4}(1-\epsilon)$	$\frac{1}{12}\epsilon$
θ_l, s_1''	$\frac{1}{12}\epsilon$	$\frac{1}{6}\epsilon$	$\frac{1}{12}\epsilon$	0

Figure 1.3: An ϵ-elaboration of $E^{*\Delta^1}$ when $\lambda = 1/2$ in Example 1.1

Here is our main question in this example: can we find a mechanism that "approximately" implements the social choice correspondence even when the designer is not certain about the information, but still "quite sure" about it? More precisely, is there a mechanism such that, for any (common prior countable) type spaces in which agents' types whose payoff type and first order belief coincide with the designer's information are sufficiently likely to happen, it implements the social choice correspondence with probability close to 1 (not necessarily with probability 1)? If there is such a mechanism, then we say F is *confidently implementable* with respect to the designer's information.

An example of such type spaces in which Δ^1 is likely to happen (here with probability $1-\epsilon$) when $\lambda = 1/2$ is depicted in Figure 1.3 . For each i and θ_i, the first order belief of type (θ_i, s_i') is $(1/2, 1/2)$, while that of (θ_i, s_i'') is different. In addition, notice that the second order belief of s_i', $\Theta_i \times \Delta(\Theta_j \times \Delta(\Theta_i))$ $j \neq i$ is different from the one when Δ^1 is common knowledge (i.e., Figure 1.2) due to the existence of (θ_j, s_j''). Note that such a type space can be substantially more complicated as we can see in "e-mail" game-like information structures (Rubinstein, 1989).

Observe that the social choice correspondence is not ex-post implementable. To see this, by the revelation principle, it is enough to check that the direct mechanism (Θ, f) where $f = F$ is not ex-post implementable. It can be easily seen by noticing that

$$u_i(f(\theta_h, \theta_l), \theta_h, \theta_l) = -1 < u_i(f(\theta_l, \theta_l), \theta_h, \theta_l) = 0.$$

Namely, in the situation where agent i has the degenerated belief that agent j's payoff type is θ_l, $j \neq i$, agent i with θ_h does not have incentive to truthfully report his type.

Nevertheless, we shall show that this social choice function is indeed confidently implementable w.r.t. the information of the designer as long as the information suggests that each agent's first order belief puts a sufficiently high probability on θ_h (i.e., λ is sufficiently large).

Note that the direct mechanism $((\Theta)_{i\in\mathcal{I}}, f)$ satisfies a "stronger" incentive compatibility if λ is sufficiently large in the sense that truth-telling is still incentive compatible *regardless of the other agents' report*, if λ is at least 1/3. To see this, first observe that the change in the probability of building the public good from one's report is independent of the opponent's report: regardless of the opponent's report, reporting θ_h increases it by 1/2. Thus, for an agent with θ_h it is incentive compatible to report θ_h regardless of the opponent's report if and only if

$$\frac{1}{2}\lambda 2 + \frac{1}{2}(1-\lambda)(-1) \geq 0 \text{ or } \lambda \geq \frac{1}{3}.$$

For an agent with θ_l, reporting θ_h is weakly dominated. We call this equilibrium $\mathbf{0}$-*dominant equilibrium* in $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$ and we call F is $\mathbf{0}$-*dominant implementable.*[11]

In a later section, we will define more generally $\mathbf{p}$-*dominant equilibrium* and $\mathbf{p}$-*dominant implementability* where $\mathbf{p} \in [0,1]^N$. Roughly, a BNE is $\mathbf{p}$-dominant if sending the equilibrium message is incentive compatible as long as the opponents send the equilibrium message with at least probability p_i, allowing sending arbitrary message with the rest probability. We will show that if there exist a mechanism and a $\mathbf{p}$-dominant equilibrium where $\sum_i p_i < 1$ that together implement the social choice correspondence in $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$, then it is confidently implementable w.r.t. the designer's first order belief information.

[11] Note that this solution concept is *different* from the standard dominant strategy equilibrium with interdependent value, since according to the definition, the incentive compatibility is required to be satisfied w.r.t. $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$; on the other hand, the standard definition of dominant strategy equilibrium requires it to be hold at any realization of agents' payoff types.

In this example, when $\lambda \geq 1/3$, (Θ, f) is $\mathbf{p}$-dominant implementable for any $\mathbf{p} \in [0,1]^2$; thus (see, e.g., Proposition 1.3), we have the following conclusion: Assume $\lambda \geq 1/3$. Then the social choice correspondence F in Example 1.1 is confidently implementable w.r.t. the designer's first order belief information.

Our main result (Theorem 1.1) substantially extends the discussion in this example; we consider the designer who has n-th order belief information about agents for some finite n. Under a condition we term *distinguishability*, we provide a sufficient condition for a social choice correspondence to be confidently implementable w.r.t. the designer's information. The sufficient condition relates $\mathbf{p}$-dominant implementability with confident implementability. We gently recommend the readers to read Section 1.5 where we elaborate further the example in this section with more detailed arguments.

1.3 Setting

1.3.1 Environment

There is a mechanism designer ("she") and finite set of agents $\mathcal{I} \equiv \{1, 2, 3, \dots, N\}$ ("he").

Let a nonempty finite set Θ_i be the set of possible *payoff types* for $i \in \mathcal{I}$. Denote by $\Theta \equiv \prod_{i \in I} \Theta_i$ the set of payoff type profiles. A payoff type of each agent represents the agent's private information about payoff-relevant parameters.

Let X be the set of *alternatives* which we assume finite. Each agent has a preference relation, which depend on $\theta \in \Theta$, over the set of lotteries $\mathcal{X} \equiv \Delta(X)$. We assume that this preference relation satisfies the conditions for having the representation for the expected utility maximization; we denote the corresponding von Neumann-Morgenstern utility by $u_i(\cdot, \theta) : X \to \mathbb{R}$ for each $i \in \mathcal{I}$ and $\theta \in \Theta$.

The designer wants to achieve a "desirable" outcome, which may depend on agents type profile, through a mechanism; and such desirability is formally modeled by a social choice correspondence. A *social choice correspondence* F is a mapping from Θ to $2^{\mathcal{X}} \setminus \{\varnothing\}$. When for each $\theta \in \Theta$, $F(\theta)$ is a singleton, we call F is a *social choice function.*

Denote $\Xi \equiv (\mathcal{I}, (\Theta_i)_{i \in \mathcal{I}}, (u_i)_{i \in \mathcal{I}}, X, F)$ and let us call this *environment* and it is common knowledge among agents and the designer.

1.3.2 Type Space and the Universal Type Space

A type space (Harsanyi, 1967) is a convenient device to model agents' payoff type and belief hierarchy.

Definition 1.1. A *type space* is defined as a tuple $\mathcal{T} = ((T_i)_i, (\tilde{\beta}_i)_i, (\tilde{\theta}_i)_i)$ where T_i is a (potentially infinite) set

$$\tilde{\beta}_i : T_i \to \Delta(T_{-i})$$

and

$$\tilde{\theta}_i : T_i \to \Theta_i.$$

The *universal type space* (Harsanyi, 1967; Mertens and Zamir, 1985; Brandenburger and Dekel, 1993) $\mathcal{T}^* = ((T_i^*)_i, (\beta_i^*)_i)$ is a type space that includes all the belief hierarchies that are *coherent*:

$$\begin{aligned} T_i^{*0} &\equiv \Theta_i \\ T_i^{*1} &\equiv \Theta_i \times \Delta(T_{-i}^{*0}) \\ &\cdots \\ T_i^{*n} &\equiv \Theta_i \times \Delta(T_{-i}^{*n-1}), \forall n \in \mathbb{N}. \end{aligned} \tag{1.1}$$

and $T_i^* \equiv \prod_{n=0}^{\infty} T_i^{*n}$, where β_i^* the Mertens and Zamir homeomorphism (Mertens and Zamir, 1985).[12]

In the conventional mechanism design theory, one usually uses a type space which is a subset of the universal type space; and choice of such a type space represents what kind of hierarchy of beliefs the designer thinks possible. Any type space $\mathcal{T} = ((T_i)_i, (\tilde{\beta}_i)_i, (\tilde{\theta}_i)_i)$ space

[12] Roughly, the homeomorphism assigns a subjective belief over other agents' belief hierarchy to a type so that the induced belief hierarchy from the subjective belief coincides with the type in the universal type space.

induce belief hierarchy of each agent. Define for each i,

$$h_i^n : T_i \to T_i^{*n}, \forall n \in \mathbb{N}$$

to be the mapping that assigns a n-th order belief to a type in the space; let $h_i(t_i) := (\tilde{\theta}_i(t_i), h_i^1(t_i), \dots) \in T_i^*$ (i.e., the entire belief hierarchy that corresponds to t_i). Also let for each $t \in T$, $h(t) := (h_i(t_i))_i$. Although $\tilde{\theta}_i$, h_i^n and h_i depend on type spaces, we shall omit such dependence for notational convenience.

Given a type space, there might be multiple types of an agent that induce the same payoff type and belief hierarchy, while it still gives different information about others' type.

Definition 1.2. A type t_i in a type space $\mathcal{T} = ((T_i)_i, (\tilde{\beta}_i)_i, (\tilde{\theta}_i)_i)$ is *redundant* if there is $t_i' \in T_i$ with $t_i' \neq t_i$ such that $h_i(t_i) = h_i(t_i')$.

It is well known that some solution concepts, for example Bayes Nash equilibrium, are affected by such redundant types, because payoff-irrelevant information could be used as a correlation device.[13]

Definition 1.3. A countable type space $\mathcal{T} = ((T_i)_i, (\tilde{\beta}_i)_i, (\tilde{\theta}_i)_i)$ allows a *common prior* if there exists a common prior $\beta \in \Delta(T)$ such that for each $i \in \mathcal{I}$ and $t_i \in T_i$,

$$\tilde{\beta}_i(t_{-i}|t_i) = \frac{\beta(t_i, t_{-i})}{\sum_{t_{-i}} \beta(t_i, t_{-i})}, \forall t_{-i} \in T_{-i}$$

for all $t_i \in T_i$ s.t. $\sum_{t_- \in T_{-i}} \beta_i(t_i, t_{-i}) > 0$.

The common prior assumption implies that any difference in posterior beliefs *only* comes from difference in information; in other words, if agents have the same information, they should have the same beliefs.[14]

Definition 1.4. Let T be a subset of T^*. T is a *belief closed subset* of T^* if

$$\forall i \in \mathcal{I}, t_i \in T_i, \beta_i^*(t_{-i}|t_i) > 0 \Rightarrow (t_i, t_{-i}) \in T. \tag{1.2}$$

[13] Some solution concepts are not affected by redundant types, e.g., interim correlated rationalizability (Dekel et al., 2007); also see Ely and Peski (2006).

[14] For further reading for the common prior assumption, refer to Aumann (1976, 1987, 1998), Gul (1998), and Morris (1995).

λ	θ'_2	θ''_2
θ'_1	$\frac{1}{4}$	$\frac{1}{4}$
θ''_1	$\frac{1}{4}$	$\frac{1}{4}$

λ'	θ'_2	θ''_2
θ'_1	$\frac{4}{9}$	$\frac{2}{9}$
θ''_1	$\frac{2}{9}$	$\frac{1}{9}$

λ' or λ''	θ'_2, s'_2	θ''_2, s'_2	θ'_2, s''_2	θ''_2, s''_2
θ'_1, s'_1	$\alpha\frac{1}{4}$	$\alpha\frac{1}{4}$	0	0
θ''_1, s'_1	$\alpha\frac{1}{4}$	$\alpha\frac{1}{4}$	0	0
θ'_1, s''_1	0	0	$(1-\alpha)\frac{4}{9}$	$(1-\alpha)\frac{2}{9}$
θ''_1, s''_1	0	0	$(1-\alpha)\frac{2}{9}$	$(1-\alpha)\frac{1}{9}$

Figure 1.4: Two minimal belief subspaces and a non minimal belief subspace that consists of the two

A belief closed subset is *minimal* if it has no proper subset that is a belief closed subset itself.

1.3.3 Mechanism and Implementation

Definition 1.5. A *mechanism (a game form)* is a pair $((M_i)_{i\in\mathcal{I}}, g)$ where M_i is a nonempty set for each $i \in \mathcal{I}$ and $g : M \to \mathcal{X}$.

We call M_i the *message space* for agent i and call g the *outcome function*. Note that (M, g) may be an extensive-form.

A particularly simple class of mechanisms is *direct* mechanisms. In a direct mechanism agents are supposed to report their type, i.e., $M_i = T_i$ for each $i \in \mathcal{I}$.

A type space $\mathcal{T}$ and a mechanism, $\mathcal{M}$ induce a Bayesian game $(\mathcal{M}, \mathcal{T})$.

Definition 1.6. Given $(\mathcal{M}, \mathcal{T})$, a strategy profile $\sigma = (\sigma_i)_{i\in\mathcal{I}}$, where $\sigma_i : T_i \to \Delta(M_i)$ be a *Bayes Nash equilibrium* (henceforth BNE) if for each $i \in \mathcal{I}$, $t_i \in T_i$ and $m_i \in M_i$ with $\sigma_i(t_i)[m_i] > 0$,

$$m_i \in \arg\max_{m'_i\in M_i} \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(m_i, \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})).$$

A mechanism (M, g) *(partially) implements in BNE* a social choice correspondence F in a common prior type space $\mathcal{T}$, if there exists a Bayes Nash equilibrium $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ such that for any $t \in T$ s.t. $\beta(t) > 0$,

$$g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

And we call such F is *(partially) implementable in BNE.* In words, the notion requires the existence of an equilibrium that yields the desirable outcome for each realization of payoff type profile.[15]

1.4 General Framework and Robustness Concept

In this section, we provide a formal framework to study the robustness notion which was informally introduced in the introduction.

1.4.1 Modeling Designer's Information

We model designer's information as a subset of the universal type space. Note that by taking this approach, we implicitly assume that the designer does not know anything about agents' payoff-irrelevant private information.

Definition 1.7. Let $E^* \subseteq T^*$ and $\epsilon > 0$. A countable common prior type space $\mathcal{T} = ((T_i)_{i \in \mathcal{I}}, \beta)$ is *ϵ-elaboration of* E^* if

$$\beta(E) \geq 1 - \epsilon$$

where $E \equiv \{t \in T : h(t) \in E^*\}$.

Denote the set of ϵ-elaboration of E^* by $\mathcal{E}(E^*, \epsilon)$.

We should emphasize that E^* is typically not belief closed. One such example is that the designer has information up to agents' possible first order beliefs. As this does not restrict agents' second order belief, agents' second order belief may put some positive probability

[15] The notion of partial implementation is different from *full implementation*, which requires *every* equilibrium to achieve the desirable outcome.

on the other agents' having some first order belief outside the designer's information. This cannot happen if E^* is belief closed; as such first order information is common knowledge. In addition, if E^* is restricted to be belief closed, then $\beta(E) \geq 1 - \epsilon$ implies that E^* is common knowledge with a high probability, which is not a part of the designer's information: the designer only has information agents' first order beliefs, but do not have information about whether it is any order of mutual knowledge, especially common knowledge. If we only consider non redundant types, then an ϵ-elaboration of E^* is any consistent belief closed subset in the universal type space in which event E^* is likely to happen. Any belief closed subset that is contained within E^* is also an elaboration; especially, in this case, E^* is common knowledge.

We also should note that the notion of ϵ-elaboration allows any beliefs and belief hierarchies of agents may happen, as long as they happen with a small probability. This concept of localness contrasts with the existing notions, e.g., Artemov et al. (2013) where some first order belief information is assumed to be commonly known to the agents and the designer; hence, in their framework, the beliefs and higher order beliefs that are inconsistent with the information cannot happen.

Lastly, let us make a remark about a subtle point: by adopting this approach, we implicitly assume that whichever a type space is true, the type space is common knowledge among the agents (but not known to the designer).[16]

1.4.2 Confident Implementation

Definition 1.8. A mechanism $\mathcal{M} = ((M_i)_{i \in \mathcal{I}}, g)$ *confidently implements* a social choice correspondence $F : \Theta \rightrightarrows \mathcal{X}$ with respect to $E^* \subseteq T^*$ if for any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for any $\epsilon \leq \bar{\epsilon}$ and for any $\mathcal{T} \in \mathcal{E}(E^*, \epsilon)$, there exists an equilibrium $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ s.t.

$$\beta(\{t \in T : g(\sigma(t)) \in F(\tilde{\theta}(t))\}) \geq 1 - \delta.$$

[16] Note that we take account potentially "large" type spaces by considering any countable type spaces. Nevertheless, we admit we do not relax common knowledge assumption among agents as we do not take into account the universal type space as it is uncountable.

In words, in any (countable and common prior) type spaces in which the event, which corresponds to the designer's information, is sufficiently likely to happen, the mechanism achieves an element of the social choice correspondence with probability arbitrarily close to 1.

1.4.2.1 Discussion: Designer's Information and Implementable Social Choice Correspondences

The situation in which the designer does not have any information about agents is captured by $E^* = T^*$. As the designer obtains more and more information, the designer is quite sure about smaller E^*. Intuitively, if the designer has more information about agents we expect that the set of confidently implementable social choice correspondences becomes larger. The following simple observation exploits this intuition.

Proposition 1.1 (Monotonicity)**.** *Suppose $E^* \subseteq E^{*'} \subseteq T^*$, if a social choice correspondence F is confidently implementable w.r.t. $E^{*'}$, then it is also confidently implementable w.r.t. E^*.*

Proof. See Appendix. □

This result also suggests why our question is meaningful given that we know ex-post implementability is confidently implementable (Bergemann and Morris, 2005). As we shall show that more social choice correspondences are confidently implementable if the designer has some information, for example, the first order belief of agents.

Corollary 1.1. *If a social choice correspondence is ex-post implementable, then it is confidently implementable w.r.t. any $E^* \subseteq T^*$.*

Note that, however, we do not know whether there is some E^* that makes the set-inclusion *strictly* hold. Indeed, we already discussed such an example in Example 1.1.

1.4.3 p-dominant Implementation

We extend **p**-dominant equilibrium (Morris et al., 1995; Kajii and Morris, 1997), which was originally defined in complete information games, to Bayesian games.[17]

Definition 1.9. Let $\mathbf{p} \in [0,1]^N$. Given a game $(\mathcal{T}, \mathcal{M})$, a strategy profile $\sigma \equiv (\sigma_i)_{i\in\mathcal{I}}$ where $\sigma_i : T_i \to \Delta(M_i)$ is a **p**-*dominant equilibrium* if for any $i \in \mathcal{I}$ with $t_i \in T_i$, $m_i \in M_i$ with $\sigma_i(t_i)[m_i] > 0$,

$$m_i \in \arg\max_{m_i' \in M_i} \sum_{t_{-i}} \beta_i(t_{-i}|t_i) u_i(g(m_i', \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})))$$

for any $\phi_{-i}(t_{-i}) \in \Delta(M_{-i})$ such that

$$\phi_{-i}(t_{-i}) = q_i^{t_{-i}} \sigma_{-i}(t_{-i}) + (1 - q_i^{t_{-i}}) \psi_{-i}(t_{-i}) \tag{1.3}$$

for some $q_i^{t_{-i}} \geq p_i$ and $\psi_{-i}(t_{-i}) \in \Delta(M_{-i})$.

In words, a strategy profile constitutes **p**-dominant equilibrium if for each agent i and θ_i, the equilibrium strategy is a best response to any conjecture over the opponents' message profile that puts on probability at least p_i on the equilibrium strategy profile; for the rest probability $1 - p_i$, the opponents' strategies are allowed to be correlated across agents; while not correlated across types within an agent. Intuitively, this solution concept captures a kind of lack of confidence of agents about the opponents' behavior. We will call ψ_j in (1.3) *babbling* of agent j.

In addition, given a mechanism, if σ is a **p**-dominant equilibrium; then it is also a $\mathbf{p}'$-dominant equilibrium for any $\mathbf{p}' \geq \mathbf{p}$.[18] In particular, any **p**-dominant equilibrium is a Bayesian Nash equilibrium. Clearly, **p**-dominant equilibrium may not exist.[19] With private value, when $\mathbf{p} = \mathbf{0}$, this notion is equivalent to (weakly) dominant strategy equilibrium. However, it turns out to be weaker than dominant strategy equilibrium with interdependent

[17] As noted in Morris et al. (1995), the notion of p-dominance is a generalization of Harsayni and Selten's risk-dominance in 2×2 games in the sense that it coincides risk dominance when $\mathbf{p} = (1/2, 1/2)$.

[18] $\mathbf{p}' \geq \mathbf{p}$ if each $p_i' \geq p_i$ for all i.

[19] In this regard, see also relevant concepts (p-BR, p-MBR) in Tercieux (2006).

value. In particular, we should emphasize that the set of $\mathbf{p}$-dominant equilibrium depends on the underlying type space $\mathcal{T}$, even when $\mathbf{p} = 0$.

Definition 1.10. A social choice correspondence $F : \Theta \rightrightarrows \mathcal{X}$ is $\mathbf{p}$-*dominant implementable* in a type space $\mathcal{T}$ if there exists a mechanism $\mathcal{M} = (M, g)$ and a $\mathbf{p}$-dominant equilibrium σ in $(\mathcal{M}, \mathcal{T})$ such that for each $t \in T$

$$g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

Note that it is a refinement of partial implementation in BNE (Definition 1.6), simply because a $\mathbf{p}$-dominant equilibrium is a Bayes Nash equilibrium.

Proposition 1.2 (Revelation principle for $\mathbf{p}$-dominant implementation)**.** *Let $\mathcal{M} = (M, g)$ be a mechanism, and let $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ where $\sigma_i : T_i \to \Delta(M_i)$ be a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}, \mathcal{T})$. Then there exists a direct mechanism $\mathcal{M}' = (T, f)$ such that*

(1) Truthful reporting, i.e., $\sigma'_i(t_i) = t_i$ for all $i \in \mathcal{I}$, is a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}', \mathcal{T})$.

(2) For every $t \in T$,

$$f(t) = g(\sigma(t)).$$

Note that if $g(\sigma(t)) \in F(\tilde{\theta}(t))$, then $f(t) \in F(\tilde{\theta}(t))$.

Due to this result, from now on we focus on direct mechanisms when we consider $\mathbf{p}$-dominant implementability.

1.5 Illustration with First Order Belief and Second Order Belief Information

1.5.1 First Order Belief Information

In this subsection, we focus on a situation where the designer has some information about the *payoff type and the first order belief* of each agent, although she is not certain about it, meaning that there is some small probability of having "unexpected" event in the true type space. This situation is modeled in our framework as follows: for each $i \in \mathcal{I}$, let Δ_i^1 be a finite subset of T_i^{*1} (refer to (1.1)) for definition) with a typical element of $(\theta_i, \lambda_i) \in T_i^{*1}$. For each $i \in \mathcal{I}$, define

$$E_i^{*\Delta_i^1} := \bigcup_{(\theta_i,\lambda_i)\in\Delta_i^1} \{t_i \in T_i^* : h_i^1(t_i) = (\theta_i, \lambda_i)\}$$

and $E^{*\Delta^1} \equiv \prod_{i\in\mathcal{I}} E_i^{*\Delta_i^1}$. Note that $E^{^*\Delta^1}$ only restricts agents' belief hierarchy up to the first order; there is no restriction beyond the first order belief.

We put restrictions on Δ_i^1: for all $(\theta_i, \lambda_i), (\theta_i', \lambda_i') \in \Delta_i^1$,

$$(\theta_i, \lambda_i) \neq (\theta_i', \lambda_i') \Rightarrow \theta_i \neq \theta_i'.$$

Put differently, $(\theta_i, \lambda_i), (\theta_i', \lambda_i') \in \Delta_i^1$,

$$\theta_i = \theta_i' \Rightarrow \lambda_i = \lambda_i'. \tag{1.4}$$

That is, the designer's information indicates that each payoff type of agent i is involved with a unique first order belief. Note that this restriction is implicitly assumed in the standard type spaces used in the mechanism design which is called by *payoff type space* by Bergemann and Morris (2005).

We also impose the following restriction on Δ_i^1 for each i: for any $(\theta_i, \lambda_i) \in \Delta_i^1$, there is some type space $\mathcal{T} = ((T_i)_i, \beta)$ such that $E^{*\Delta^1}$ happens with probability 1 (i.e., $\beta(\{t \in T : h(t) \in E^{*\Delta^1}\}) = 1$), and in this type space (θ_i, λ_i) happens with positive probability, i.e.,

$$\exists t_i \in T_i \text{ s.t. } h_i^1(t_i) = (\theta_i, \lambda_i). \tag{1.5}$$

Note that these conditions are satisfied Δ_i^1 in Example 1.1.

The situation that the designer is "quite sure" about this event, but not necessarily certain, is formally described in our framework as follows: whenever the true type space is $\mathcal{T} = ((T_i)_i, \beta)$

$$\beta(\{t \in T : h(t) \in E^{*\Delta^1}\}) \geq 1 - \epsilon.$$

We call such a countable and common prior type space an ϵ-elaboration of $E^{*\Delta^1}$.

Given a social choice correspondence, the designer wants to make a mechanism that "approximately" implements the social choice correspondence in any ϵ-elaboration of $E^{*\Delta^1}$ for small ϵ; this question is formally described in our framework as follows: is there any mechanism $\mathcal{M} \equiv ((M_i)_i, g)$ such that for any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for any $\epsilon < \bar{\epsilon}$, for any ϵ-elaboration of $E^{*\Delta^1}$, there exists an equilibrium $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ such that

$$\beta(\{t \in T : g(\sigma(t)) \in F(\tilde{\theta}(t))\}) \geq 1 - \delta. \tag{1.6}$$

That is, by informally saying the mechanism "approximately" achieves the social choice correspondence we mean that the social choice correspondence is achieved with probably *arbitrarily* close to 1, although not necessary 1, in *any* ϵ-elaboration for sufficiently small ϵ. As defined previously, if there is such a mechanism, we say that the social choice correspondence is *confidently implementable w.r.t.* $E^{*\Delta^1}$.

How can we identify such a mechanism? Consider the situation where $E^{*\Delta^1}$ is *certain* to the designer. This means that the designer is certain that the true type space (given there is no redundant types) is included in this event, i.e., the true type space corresponds to a (consistent) *belief closed subset* in this event. Note that $E^{*\Delta^1}$ is common knowledge in a belief closed subset.

Let us now examine the implication of condition (1.4). Note that for each payoff type of agent i, there is a unique first order belief based on the designers' information, and it is common knowledge among agents in a belief closed subset; this implies that the second order and higher order beliefs are also determined *only* by agents' payoff type. Thus, we may describe the common prior equivalently only using each agent payoff types, i.e., $\hat{\lambda} \in \Delta(\Theta)$. Note that this type space $((\Theta_i)_{i \in \mathcal{I}}, \hat{\lambda})$ satisfies exactly the definition of a payoff type space. *Suppose* that there is a unique belief closed subset in $E^{*\Delta^1}$ (which is the case in Example 1.1). Then, by a well-known result, there is a *unique* common prior that is consistent with agents' belief hierarchies (see Figure 1.2 for Example 1.1). Our main result, which will be provided in a following section, does *not* involve the uniqueness of belief closed subset; while it *does* involve a generalization of condition (1.4).

At this point, let us provide an overview of our argument for finding a mechanism that confidently implements a social choice correspondence F w.r.t. $E^{*\Delta^1}$. We first find a (direct) mechanism and a BNE $\sigma \equiv (\sigma_i)_{i\in\mathcal{I}}$ that achieves F in the belief closed subset or equivalently $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$. Then, we show that in any ϵ-elaboration of $E^{*\Delta^1}$, we can find an equilibrium $\sigma' \equiv (\sigma'_i)_{i\in\mathcal{I}}$ in this ϵ-elaboration in which type profiles in a certain event, which will be described carefully shortly, play the same action as in σ. Then, we will show that the probability of the event becomes close to 1 as ϵ goes to 0, thereby implying that σ' achieves F with probability close to 1, i.e., satisfies condition (1.6).

To be more specific, consider the direct mechanism $((\Theta_i)_{i\in\mathcal{I}}, f)$ in $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$, i.e., $f : \Theta \to \mathcal{X}$; and suppose it allows the truth-telling BNE $\sigma \equiv (\sigma_i)_{i\in\mathcal{I}}$, where $\sigma_i(\theta_i) = \theta_i$ for all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, that satisfies a "stronger" incentive compatibility in the following sense: for a given $\mathbf{p} \in [0,1]^N$, for each agent i,

$$\theta_i \in \arg\max_{\theta'_i\in\Theta_i} \sum_{\theta_{-i}\in\Theta_{-i}} \hat{\lambda}_i(\theta_{-i}|\theta_i) u_i(f(\theta'_i, \phi_{-i}(\theta_{-i})), \theta_i, \theta_{-i}) \tag{1.7}$$

for any $\phi_{-i}(\theta_{-i}) = q_i\sigma_{-i}(\theta_{-i}) + (1-q_i)\psi_{-i}(\theta_{-i})$ where $\psi_{-i}(\theta_{-i}) \in \Delta(\Theta_{-i})$ and $q_i \geq p_i$. In words, agent i's equilibrium action (i.e., truth-telling) is a best response to any conjecture over others' report that satisfies the condition that truthful reporting happens with at least probability p_i. We call this particular BNE $\mathbf{p}$-*dominant equilibrium* in type space $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$.

Let $\mathcal{T} = ((T_i)_i, \beta)$ be an ϵ-elaboration. For notational convenience, let $E_i^{\theta_i} \equiv \{t_i \in T_i : \tilde{\theta}_i(t_i) = \theta_i\}$; $E^{\theta} \equiv \prod_{i\in\mathcal{I}} E_i^{\theta_i}$; and also let $E_i^{\Delta_i^1} \equiv \{t_i \in T_i : h_i(t_i) \in E_i^{*\Delta_i^1}\}$. Consider a strategy profile in $\mathcal{T}$, $\sigma' \equiv (\sigma'_i)_{i\in\mathcal{I}}$, where $\sigma'_i : T_i \to \Delta(\Theta_i)$, such that for each $\theta \in \Theta$, and $t \in C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})$

$$\sigma'_i(t_i) = \sigma_i(\theta_i) = \theta_i$$

where $C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta}) \subseteq E^{\Delta^1} \cap E^{\theta}$ is the event in which every agent i believes E^{Δ^1} with probability at least p_i; and believes that every agent j, $j \neq i$, believes E^{Δ^1} with probability at least p_j, with probability at least p_i and so on, *conditional on* E^{θ}. Formally, given $\mathbf{p} \in [0,1]^N$,

$$B_i^{p_i}(E^{\Delta^1}|E^{\theta}) := \{t_i \in T_i : t_i \in E_i^{\Delta_i^1} \cap E_i^{\theta_i} \text{ and } \beta_i(E_{-i}^{\Delta_{-i}^1}|t_i, E_{-i}^{\theta_{-i}}) \geq p_i\}$$

$$B_*^{\mathbf{p}}(E^{\Delta^1}|E^{\theta}) := \prod_{i\in\mathcal{I}} B_i^{p_i}(E^{\Delta^1}|E^{\theta})$$

and

$$C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta}) := \bigcap_{n\geq 1}[B_*^{\mathbf{p}}]^n(E^{\Delta^1}|E^{\theta}).$$

Note that we have not yet specified strategy profile for types outside $C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})$ for some θ. For the time being, assume that agent with such type reports some arbitrary payoff type. Let $t \in C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})$ and agent i with type t_i s.t.

$$h_i^1(t_i) = (\theta_i, \hat{\lambda}_i(\cdot|\theta_i)),$$

we can find such θ_i in $((\Theta_i)_i, \hat{\lambda})$ by condition (1.5). The relevant incentive compatibility is then

$$\sum_{\theta_{-i}} \hat{\lambda}_i(\theta_{-i}|\theta_i) \sum_{t_{-i}\in C_{-i}^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})} \beta_i(t_{-i}|t_i, \tilde{\theta}_{-i}(t_{-i}) = \theta_{-i}) u_i(f(\theta_i, \theta_{-i}), \theta_i, \theta_{-i})$$
$$+ \sum_{\theta_{-i}} \hat{\lambda}_i(\theta_{-i}|\theta_i) \sum_{t_{-i}\in T_{-i}\setminus C_{-i}^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})} \beta_i(t_{-i}|t_i, \tilde{\theta}_{-i}(t_{-i}) = \theta_{-i}) u_i(f(\theta_i, \sigma_i'(t_{-i}), \theta_i, \theta_{-i}).$$

But, as $t \in C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})$

$$\beta_i(C_{-i}^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})|t_i, E_{-i}^{\theta_{-i}}) \geq p_i.$$

By construction of σ', this implies that agent i believes that for each θ_{-i} the opponents play the **p**-dominant equilibrium strategy $\sigma_{-i}(\theta_{-i}) = \theta_{-i}$ with at least probability p_i. Thus, by definition of **p**-dominant equilibrium, reporting θ_i is incentive compatible, no matter what $\sigma_i'(t_{-i})$ is for $t_{-i} \notin C_{-i}^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})$ for some θ_{-i}. With slightly more complicated argument, we can actually show that there is indeed a BNE $\sigma' = (\sigma_i')_{i\in\mathcal{I}}$, $\sigma_i' : T_i \to \Delta(\Theta_i)$ in which for each $t \in C^{\mathbf{p}}(E^{\Delta^1}|F^{\theta})$, each agent i

$$\sigma_i'(t_i) = \sigma_i(\theta_i) = \theta_i.$$

What is remaining for our result is to show that the probability of the event $\beta(C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})|E^{\theta})$ is sufficiently close to 1 for each $\theta \in \text{supp}(\hat{\lambda})$. We can show that as ϵ goes to 0,

$$\beta(E^{\theta}) \to \hat{\lambda}(\theta).$$

This implies that for any $\theta \in \text{supp}(\hat{\lambda})$, and for any ϵ',

$$\beta(E^{\Delta^1}|E^{\theta}) \geq 1 - \epsilon'$$

for sufficiently small ϵ. Then, due to the *critical path* lemma (Kajii and Morris, 1997), we know that if $\mathbf{p} \in [0,1]^N$ satisfies $\sum_{i\in\mathcal{I}} p_i < 1$, then for any $\delta' > 0$,

$$\beta(C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})|E^{\theta}) \geq 1-\delta'$$

if ϵ' is sufficiently small (thus if ϵ is sufficiently small). As a result,

$$\beta(\{t \in T : f(\sigma'(t)) \in F(\tilde{\theta}(t))\}|E^{\theta}) \geq \beta(C^{\mathbf{p}}(E^{\Delta^1}|E^{\theta})|E^{\theta}) \geq 1-\delta'$$

and thus

$$\beta(\{t \in T : f(\sigma'(t)) \in F(\tilde{\theta}(t))\}) \geq 1-\delta.$$

Thus, we have the following result:

Proposition 1.3. *Let $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$ be the common prior type space corresponding to the minimal consistent belief closed subset in $E^{*\Delta^1}$. If a social choice correspondence F is* $\mathbf{p}$-*dominant implementable in $((\Theta_i)_{i\in\mathcal{I}}, \hat{\lambda})$ for some $\mathbf{p} \in [0,1]^N$ s.t. $\sum_{i\in\mathcal{I}} p_i < 1$, then it is confidently implementable w.r.t. $E^{*\Delta^1}$.*

We shall provide a generalization of this result (Theorem 1.1) and the complete proof in the following sections.

1.5.2 Second Order Belief Information

Now we consider a situation where the designer is quite sure about agents' second order belief. Let for each agent i

$$\Delta_i^2 \subseteq T_i^{*2} = \Theta_i \times \Delta(T_{-i}^{*1}) \subseteq \Theta_i \times \prod_{j\neq i} \Delta(\Theta_j \times \Delta(\Theta_{-j}))$$

and be finite. Let $\Delta^2 \equiv (\Delta_i^2)_{i\in\mathcal{I}}$. Given Δ^2,

$$E_i^{*\Delta_i^2} := \bigcup_{(\theta_i,\delta_i^2)\in\Delta_i^2} \{t_i \in T_i^* : h_i^2(t_i) = (\theta_i, \delta_i^2)\}$$

and

$$E^{*\Delta^2} \equiv \prod_{i\in\mathcal{I}} E_i^{*\Delta_i^2}.$$

That is, $E^{*\Delta^2}$ is the set of type profiles in the universal type space which are characterized by a certain set of the payoff types and second order beliefs. Note that it does not say anything about belief hierarchies higher than the second order. The set is interpreted as the designer has information about agents' payoff and up to the second order beliefs.

Since the second order belief also includes the first order belief this means that the designer now confidently implements weakly larger set of social choice functions (see Proposition 1.1).

Suppose that the designer's information indicates that for each agent, each possible payoff type and second order belief has different payoff type. In this case, we can apply the argument which we used for the first order belief information. Thus, we know that a social choice correspondence that is **p**-dominant implementable in E^{Δ^1} where λ corresponds to Δ^2 (i.e., the marginal of δ_i^2 on Θ_{-i} coincides with some λ_i) is confidently implementable.

Now in the designer's information represented by $E^{*\Delta^2}$, suppose that for some agent there are some pairs of payoff type and second order belief whose payoff types are the same. For example, Δ_i^2 includes

$$\delta_i^2 \equiv \left(\theta_h, \frac{1}{2}\left(\theta_h, \frac{1}{2}\theta_h \oplus \frac{1}{2}\theta_l\right) \oplus \frac{1}{2}\left(\theta_l, \theta_h \oplus \frac{1}{2}\theta_l\right)\right)$$

and

$$\delta_i'^2 \equiv \left(\theta_h, \frac{2}{3}\left(\theta_h, \frac{2}{3}\theta_h \oplus \frac{1}{3}\theta_l\right) \oplus \frac{1}{3}\left(\theta_l, \frac{1}{3}\theta_h + \frac{2}{3}\theta_l\right)\right).$$

In this case, we cannot apply the argument that was employed for the first order beliefs. In particular, it is not anymore the case that the strategy of each agent can be regarded determined only by payoff types. How can we approach then?

Here is a possible solution. Suppose that there are no pairs in Δ_i^2 which have the same (induced) *first order beliefs*. We can show that in any belief closed subset, each type of agent i has different first order belief (see Lemma 1.1 and the relevant discussion around the lemma). This implies that as long as agents identify the first order belief of an agent in the belief closed subset, they can identify his strategy.

Now consider any elaboration of $E^{*\Delta^2}$, $\mathcal{T} = ((T_i)_i, \beta)$; consider a strategy profile in which if type profile $t \in C^{\mathbf{p}}(E^{\Delta^2}|E^{(\theta,\lambda)})$, where

$$E^{\Delta^2} \equiv \{t \in T : h(t) \in E^{*\Delta^2}\}$$
$$E_i^{(\theta_i,\lambda_i)} \equiv \{t_i \in T_i : h_i^1(t_i) = (\theta_i, \lambda_i)\}$$

and

$$E^{(\theta,\lambda)} \equiv \prod_{i \in \mathcal{I}} E_i^{(\theta_i,\lambda_i)},$$

then each agent i employs the same strategy of the type at in the maximal belief closed subset who has the same *first order belief* (or equivalently the same second order belief because such different first order belief implies different second order belief), i.e., if

$$t \in C^{\mathbf{p}}(E^{\Delta^2}|E^{(\theta,\lambda)})$$

and

$$h_i^1(t_i) = h^1(\hat{t}_i)$$

then agent i plays $\sigma_i'(t_i) = \sigma_i(\hat{t}_i)$. Let us check this type's incentive compatibility:

$$\begin{aligned}
&\sum_{t_{-i} \in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(m_i', \sigma_{-i}'(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t_{-i})) \\
&= \sum_{(\theta_{-i},\lambda_{-i}) \in T_{-i}^{*1}} \sum_{t_{-i}: h_{-i}^1(t_{-i}) = (\theta_{-i},\lambda_{-i})} \beta_i(t_{-i}|t_i) u_i(g(m_i', \sigma_{-i}'(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t_{-i})) \\
&= \sum_{(\theta_{-i},\lambda_{-i}) \in T_{-i}^{*1}} \delta_i^2(\theta_{-i}, \lambda_{-i}) \sum_{t_{-i}: h_{-i}^1(t_{-i}) = (\theta_{-i},\lambda_{-i})} \beta_i(t_{-i}|t_i, \theta_{-i}, \lambda_{-i}) u_i(g(m_i', \sigma_{-i}'(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t_{-i}))
\end{aligned}$$

Suppose $(\theta_i, \delta_i^2) = h_i^2(\hat{t}_i)$ for some $\hat{t}_i \in \hat{T}_i$. Then, $\delta_i^2 \in \Delta(T_{-i}^{*1})$ only has the positive density on $(\theta_{-i}, \lambda_{-i})$ such that there exists $\hat{t}_{-i}$ such that $h_{-i}^1(\hat{t}_{-i}) = (\theta_{-i}, \lambda_{-i})$. In addition, since every $\hat{t}_j$ has different first order belief, each first order belief on the support is matched *only* *one* $\hat{t}_j$ for each $j \neq i$. Thus, we may rewrite the expression as

$$\sum_{\hat{t}_{-i} \in \hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i) \sum_{t_{-i}: h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})) u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))$$

$$= \sum_{\hat{t}_{-i} \in \hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i)$$

$$\times \Bigg(\sum_{t_{-i} \in C^{\mathbf{p}}_{-i}(E^{\Delta^2}|E^{(\theta,\lambda)}): h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})) u_i(g(m'_i, \sigma_{-i}(\hat{t}_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))$$

$$+ \sum_{t_{-i} \in T_{-i} \setminus C^{\mathbf{p}}_{-i}(E^{\Delta^2}|E^{(\theta,\lambda)}): h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^1_{-i}(t_{-i}) = h^1_{-i}(\hat{t}_{-i})) u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i})) \Bigg)$$

The rest argument is similar to the first order belief case. In addition, this argument will be generalized in the following sections.

1.6 Confident Implementation w.r.t. n-th Order Belief Information

Let us consider the situation where the designer has n-th order belief information about each agent for some finite n.

Definition 1.11. An *n-th order belief event for agent i* is defined by

$$E_i^{*\Delta_i^n} := \bigcup_{(\theta_i, \delta_i^n) \in \Delta_i^n} \{t_i \in T_i^* : h_i^n(t_i) = (\theta_i, \delta_i^n)\}$$

where Δ_i^n is a finite subset of T_i^{*n}. Define *n-th order belief event* by

$$E^{*\Delta^n} \equiv \prod_{i \in \mathcal{I}} E_i^{*\Delta_i^n}$$

where $\Delta^n \equiv (\Delta_i^n)_i$ and for each i, Δ_i^n is a finite subset of T_i^{*n}.

Definition 1.12. n-th order belief event $E^{*\Delta^n}$ is *regular* if

(i) It has a nonempty consistent belief closed subset.

(ii) For any agent i and $t_i \in E_i^{*\Delta_i^n}$, there exists a consistent belief closed subset in $E^{*\Delta^n}$, $((\hat{T}_i)_{i \in \mathcal{I}}, \hat{\beta})$ and $\hat{t}_i \in \hat{T}_i$ with $\sum_{\hat{t}_{-i} \in \hat{T}_{-i}} \hat{\beta}(\hat{t}_i, \hat{t}_{-i}) > 0$ such that $h_i^n(t_i) = h_i^n(\hat{t}_i)$.

The first condition requires that if the designer's information is correct, there exists a common prior type space that is consistent to the information. The second condition says that the designer's information should be "consistent" in the sense that if the designer's information is correct, there should be a consistent belief closed subset in which there is a type whose n-th order belief coincides the designer's information.

Assumption 1.1 (Regularity). *n-th order belief event $E^{*\Delta^n}$ is regular.*

We maintain this assumption throughout this section. To motivate this assumption, let us consider the following example.

Example 1.2. Suppose that there are two agents $\mathcal{I} = \{1, 2\}$ with two payoff types $\Theta_i \equiv \{\theta_h, \theta_l\}$ for each agent i. Consider the following designer's information about agents' first order belief:

$$\Delta_1^1 = \left\{ \left(\theta_h, \left(\frac{2}{3}, \frac{1}{3}\right)\right), \left(\theta_l, \left(\frac{1}{3}, \frac{2}{3}\right)\right) \right\}$$
$$\Delta_2^1 = \left\{ \left(\theta_h, \left(\frac{1}{2}, \frac{1}{2}\right)\right), \left(\theta_l, \left(\frac{1}{2}, \frac{1}{2}\right)\right) \right\}.$$

We claim that there is no consistent belief closed subset in the corresponding $E^{*\Delta^1}$ where $\Delta^1 \equiv \Delta_1^1 \times \Delta_2^1$. As each payoff type of each agent corresponds only one belief hierarchy, we may focus on a prior on Θ. Suppose that there is a common prior $p \in \Delta(\Theta)$ that induces such beliefs. Let $\alpha \equiv p(\theta_h, \theta_h) + p(\theta_h, \theta_l)$. In order to have the first order belief of agent 2 with θ_h, $\left(\frac{1}{2}, \frac{1}{2}\right)$,

$$\frac{\alpha \frac{2}{3}}{\alpha \frac{2}{3} + (1-\alpha)\frac{1}{3}} = \frac{1}{2}$$

Then, $\alpha = \frac{1}{3} < 1/2$. But, then it violates the first belief of agent 2 with θ_l,

$$\frac{\alpha \frac{1}{3}}{\alpha \frac{1}{3} + (1-\alpha)\frac{2}{3}} = \frac{1}{5} \neq \frac{1}{2}.$$

Hence, a contradiction. This is related to the first condition for the regularity. Now, consider

$$\tilde{\Delta}_1^1 = \left\{ \left(\theta_h, \left(\frac{2}{3}, \frac{1}{3}\right)\right), \left(\theta_l, \left(\frac{1}{3}, \frac{2}{3}\right)\right), \left(\theta_h, \left(\frac{1}{2}, \frac{1}{2}\right)\right), \left(\theta_l, \left(\frac{1}{2}, \frac{1}{2}\right)\right) \right\}$$

and let $\tilde{\Delta}_2^1 = \Delta_2^1$. In this case, we can easily see that there is a nonempty consistent belief closed subset in $E^{*\tilde{\Delta}^1}$ with a prior for which each payoff type is independently and identically drawn for each agent with equal probability. However, some information about agent 1, $\left(\theta_h, \left(\frac{2}{3}, \frac{1}{3}\right)\right), \left(\theta_l, \left(\frac{1}{3}, \frac{2}{3}\right)\right)$, cannot be possible in a consistent belief closed subset; thus $E^{*\tilde{\Delta}^n}$ is still not regular by the second condition for the regularity.

1.6.1 Distinguishability by k-th Order Belief

Definition 1.13. An n-th order belief event for agent i $E_i^{*\Delta_i^n}$ is *distinguishable by k-th order belief* for some $k \leq n$ if for each $t_i, t_i' \in E_i^{*\Delta_i^n}$,

$$h_i^k(t_i) = h_i^k(t_i') \Rightarrow h_i^n(t_i) = h_i^n(t_i').$$

Given n-th order belief event $E^{*\Delta^n} \equiv E_i^{*\Delta_i^n}$, if for each agent i, $E_i^{*\Delta_i^n}$ is distinguishable by k-th order belief, then we call $E^{*\Delta^n}$ is *distinguishable by k-th order belief.*

In words, the designer's information about agent i, Δ_i^n, indicates that each possible n-th order belief (this includes lower order beliefs) is distinguishable by a lower k-th order belief.

In the first order belief event case, if there are two possible first order beliefs that have the same payoff type, then it is *not* distinguishable by 0-order belief (i.e., payoff types). In this case, it is only distinguishable by the first order beliefs.

Note that if the designer's information is distinguishable by k-th order belief where $k < n$, then it is also distinguishable by k'-th order belief for any $k \leq k' \leq n$. This is simply because for $t_i, t_i' \in E_i^{*\Delta_i^n}$ if $h_i^{k+1}(t_i) = h_i^{k+1}(t_i')$ then $h_i^k(t_i) = h_i^k(t_i')$.

The next observation relates k-th order belief distinguishability of designer's information to the properties of the possible type spaces (in other words, possible belief hierarchies) when the designer is certain about her information: Suppose that the designer is certain of her information which satisfies the distinguishability condition; then this information is *common knowledge.* Based on the designer's information, for each agent i and k-th order belief, there is a unique $(k+1)$-th order belief. In this sense, his $(k+1)$-th order belief is determined by his k-th order belief. Consider his $(k+2)$-th order belief, which is a distribution over $(k+1)$-th order beliefs. Since $(k+1)$-th order belief of $j \neq i$ is determined by their k-th order belief and it is known to every agent, agent i's $(k+2)$-th order belief is essentially the same to his $(k+1)$-th order belief. Thus, $(k+2)$-th order belief is also determined by k-th order belief. This argument goes on infinitely. That is, agent i's entire belief hierarchy is determined (or summarized) by his k-th order belief.[20] Thus, in a belief closed subset each agent's type (i.e., payoff type and belief hierarchy) can be identified by payoff type and k-th order belief. As a result, different types of each agent should have different k-th order belief.

Lemma 1.1. *Suppose n-th order belief event $E^{*\Delta^n}$ is distinguishable by k-th order belief where $k < n$. Then,*

(1) In each belief closed subset in $E^{*\Delta^n}$, $\hat{T} = \prod_{i \in \mathcal{I}} \hat{T}_i \subseteq E^{*\Delta^n}$, each agent i's type has different k-th order belief, i.e., for all $\hat{t}_i, \hat{t}'_i \in \hat{T}_i$,

$$\hat{t}_i \neq \hat{t}'_i \Rightarrow h^k_i(\hat{t}_i) \neq h^k_i(\hat{t}'_i).$$

(2) Every belief closed subset in $E^{*\Delta^n}$ is finite. In addition, the number of the belief closed subsets in $E^{*\Delta^n}$ is finite.

Proof. The first item of the lemma is immediate from the discussion right before the lemma. The second item comes from the fact that Δ^n_i is finite. □

We should emphasize that the above discussion is only when the designer is certain about her information.

1.6.2 Common p-belief

Let $\mathcal{T} = ((T_i)_i, \beta)$ be a type space. An event is $E \subseteq T$ is *simple* if $E = \prod_{i \in \mathcal{I}} E_i$ for some $E_i \subseteq T_i$ for each $i \in \mathcal{I}$. Given a simple event $E \subseteq T$ and $F \subseteq T$, we straightforwardly extend the belief operator (Monderer and Samet, 1989) in a way to allow conditioning on F.[21]

$$B^{p_i}_i(E|F) := \{t_i \in T_i : t_i \in E_i \cap F_i \text{ and } \beta_i(E_{-i}|t_i, F_{-i}) \geq p_i\}.$$

[20] Note that payoff type spaces satisfy 0-th order distinguishability. In this sense, this condition generalize payoff type spaces, although the condition applies to any subsets in the universal type space.

[21] Kajii and Morris (1997) extends it by allowing asymmetric $\mathbf{p}$.

Let $B_*^{\mathbf{p}}$ be the set of states in which every i believes event E with probability at least p_i, i.e.,

$$B_*^{\mathbf{p}}(E|F) := \prod_{i \in \mathcal{I}} B_i^{p_i}(E|F).$$

Let $C^{\mathbf{p}}(E|F)$ be the set of states in which E is *common* $\mathbf{p}$*-belief conditional on* F, i.e.,

$$C^{\mathbf{p}}(E|F) := \bigcap_{n \geq 1} [B_*^{\mathbf{p}}]^n(E|F)$$

and event E is common $\mathbf{p}$-belief conditional on F at $t \in T$ if $t \in C^{\mathbf{p}}(E|F)$.

An event E is $\mathbf{p}$*-evident conditional on* F if it is $\mathbf{p}$-believed whenever it is true, i.e.,

$$E \cap F \subseteq B_*^{\mathbf{p}}(E|F).$$

Lemma 1.2. *An event E is common $\mathbf{p}$-belief conditional on F at $t \in T$ if and only if there exists $\mathbf{p}$-evident event E' conditional on F such that $t \in E' \subseteq B_*^{\mathbf{p}}(E|F)$.*

1.6.3 Common p-belief and p-dominant Implementation

Definition 1.14. A belief closed subset $\hat{T}$ in $E^* \subseteq T^*$ is *maximal* if it is the union of every belief closed subset in E^*, i.e.,

$$\hat{T} := \bigcup_{\alpha \in \mathcal{A}} \hat{T}^{\alpha}$$

where $\hat{T}^{\alpha}$ is a belief closed subset in E^*. A belief closed subset $\hat{T}$ is the *maximal consistent belief closed subset* if it is the union of every consistent belief closed subsets in E^*.

By the regulaity assumption, there is always non empty maximal consistent belief subset.

Proposition 1.4. *Let $E^{*\Delta^n}$ be a n-th order belief event for some $n \geq 1$ and it is distinguishable by k-th order belief for some $k < n$. Let $\mathcal{M} = (M, g)$ be a mechanism. Suppose $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ is a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}, \hat{\mathcal{T}})$ for some $\mathbf{p} \in [0,1]^N$ where $\hat{\mathcal{T}}$ is the type space that corresponds to the maximal belief closed subset.*

Consider a countable type space $\mathcal{T} = ((T_i)_i, \beta)$ s.t. there exists $E' \subseteq T$ s.t. $h(E') \subseteq E^{*\Delta^n}$ and $\mathbf{p}$-evident conditional on $E^k(\hat{t}) \equiv \{t \in T : h^k(t) = h^k(\hat{t})\}$ for each $\hat{t} \in \hat{T}$. Then, there *exists a BNE $\sigma' = (\sigma_i')_i$, where $\sigma_i' : T_i \to \Delta(M_i)$ in $(\mathcal{M}, \mathcal{T})$ such that for all $t \in E'$ and $i \in \mathcal{I}$,*

$$h_i^k(t_i) = h_i^k(\hat{t}_i) \Rightarrow \sigma_i'(t_i) = \sigma_i(\hat{t}_i).$$

Let us briefly explain the intuition behind this result.[22] We are claiming the existence of a particular equilibrium in any countable type space (note that this result does not involve common prior assumption; so $\mathcal{T}$ in the statement does not need to be an ϵ-elaboration.) in which the equilibrium strategy coincides with the **p**-dominant equilibrium for type profiles that are included in the event that agents n-th order belief coincides with the designer's information is **p**-evident, i.e., the set of type profiles (state of the world) that each agent believes that the opponents' n-th order belief is the same as the designer's information at least probability p_i whenever it is true. A stronger incentive compatibility that is required by the **p**-dominant equilibrium is exploited to satisfy the incentive compatibility of such type profile.

Let us also explain how the condition of distinguishability by k-th order belief is exploited. By Lemma 1.1, if the designer's information satisfies this condition, in any belief closed subsets, in particular, in the maximal belief closed subset $\hat{T}$, any type of agent i has different k-th order belief (and so different $k' > k$-th order belief). This implies that the equilibrium strategy of agent i in this type space can be identified by agent i's k-th order belief (measurable by k-th order belief). Thus, it is just enough to have "right" $(k+1)$-th order belief to make "right" conjecture about the opponents' play. As $n \geq k+1$, any agent with n-th order belief can do this. Our "approximation" of **p**-dominant equilibrium (thus, approximation of the social choice correspondence of interest) in any type space exploits this observation. Note that we have not discussed that how likely the **p**-evident event happens. This will be discussed in the following subsection shortly.

Lastly, we shall explain why we consider the maximal belief closed subset. When we consider the social choice *function*, it is true that the social choice function is **p**-dominant

[22] A similar argument is exploited in the robust prediction literature, e.g., Monderer and Samet (1989) and Kajii and Morris (1997). We will explain what is added more in our argument.

implementable in each minimal belief closed subset if and only if it is implementable in any belief closed subset; especially the maximal one. However, we conjecture that this is not the case for social choice *correspondences*. The thing is when a social choice correspondence is implemented in each minimal belief closed subset, they might achieve different selection of the social choice correspondence. Thus the direct mechanism induced may be different. On the other hand, if a social choice correspondence is **p**-dominant implementable in the maximal belief closed subset, then it obviously implies that it is **p**-dominant implementable in any belief closed subset.

Proof. To construct $\sigma' = (\sigma'_i)_{i\in\mathcal{I}}$, first consider a modified game in which each agent i with $t_i \in E'_i$ such that $h^k_i(t_i) = h^k_i(\hat{t}_i)$ is fixed to play $\sigma_i(\hat{t}_i)$. Then, from a well-known result for countable type spaces, there exits a BNE, $\sigma'' = (\sigma''_i)_{i\in\mathcal{I}}$ where $\sigma''_i : T_i \setminus E'_i \to \Delta(M_i)$. Now define $\sigma'_i : T_i \to \Delta(M_i)$ in the original game as follows:

$$\sigma'_i(t_i) = \begin{cases} \sigma_i(\hat{t}_i) & \text{if } t_i \in E'_i \text{ and } h^k_i(t_i) = h^k_i(\hat{t}_i), \\ \sigma''_i(t_i) & \text{if } t_i \in T_i \setminus E'_i. \end{cases}$$

Note that for $t_i \in E'_i$, this strategy is well-defined, because there is at most one $\hat{t}_i$ by (1) of Lemma 1.1; and there is at least one such $\hat{t}_i$ by the regularity assumption (Assumption 1.1).

Let us consider the incentive compatibility of agent i. If $t_i \in T_i \setminus E'_i$, the incentive compatibility is satisfied by construction. To check incentive compatibility for $t_i \in E'_i$, note that

$$\begin{aligned} &\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t_{-i})) \\ &= \sum_{t^{*k}_{-i}\in T^{*k}_{-i}} \sum_{t_{-i}:h^k_{-i}(t_{-i})=t^{*k}_{-i}} \beta_i(t_{-i}|t_i)u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t_{-i})) \\ &= \sum_{t^{*k}_{-i}\in T^{*k}_{-i}} \delta^{k+1}_i(t^{*k}_{-i}) \sum_{t_{-i}:h^k_{-i}(t_{-i})=t^{*k}_{-i}} \beta_i(t_{-i}|t_i, \theta_{-i}, \lambda_{-i})u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(t^{*k}_{-i})) \end{aligned}$$

(For notation T^{*k}_{-i} refer to (1.1)). Suppose $\delta^{k+1}_i(\cdot) = h^{k+1}_i(\hat{t}_i)$ for some $\hat{t}_i \in \hat{T}_i$. Then, δ_i only has the positive density for t^{*k}_{-i} such that there exists $\hat{t}_{-i}$ such that $h^k_{-i}(\hat{t}_{-i}) = t^{*k}_{-i}$. In addition, since every $\hat{t}_j$ has different k-th order belief by distinguishability, each n-th order belief on the support is matched *only one* $\hat{t}_j$ for each $j \neq i$. Thus, we may rewrite the expression as

$$\sum_{\hat{t}_{-i}\in\hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i) \sum_{t_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))$$

$$= \sum_{\hat{t}_{-i}\in\hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i)\Bigg(\sum_{t_{-i}\in E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))$$

$$+ \sum_{t_{-i}\in T_{-i}\setminus E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))u_i(g(m'_i, \sigma'_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))\Bigg)$$

$$= \sum_{\hat{t}_{-i}\in\hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i)\Bigg(\sum_{t_{-i}\in E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))u_i(g(m'_i, \sigma_{-i}(\hat{t}_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))$$

$$+ \sum_{t_{-i}\in T_{-i}\setminus E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))u_i(g(m'_i, \sigma''_{-i}(t_{-i})), \theta_i, \tilde{\theta}_{-i}(\hat{t}_{-i}))\Bigg)$$

Since $t_i \in E'_i$, note that and E^* is conditional $\mathbf{p}$-evident given $F(\hat{t})$ for each $\hat{t} \in \hat{T}$,

$$q_i^{t_i} \equiv \sum_{t_{-i}\in E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i})) \geq p_i.$$

Define

$$\phi_{-i}(t_{-i}) := q_i^{t_i}\sigma_i(\hat{t}_{-i}) + (1 - q_i^{t_i})\psi_{-i}(\hat{t}_{-i})$$

where $\psi_{-i} : \hat{T}_{-i} \to \Delta(M_{-i})$,

$$\psi_{-i}(\hat{t}_{-i}) := \frac{1}{1 - q_i^{t_i}} \sum_{t_{-i}\in T_{-i}\setminus E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i}))\sigma''_{-i}(\hat{t}_{-i}, s_{-i}).$$

Note that

$$\sum_{m_{-i}} \psi_{-i}(\hat{t}_{-i})[m_{-i}]$$

$$= \frac{1}{1 - q_i^{t_i}} \sum_{t_{-i}\in T_{-i}\setminus E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i})) \sum_{m_{-i}} \sigma''_{-i}(\hat{t}_{-i}, s_{-i})[m_{-i}]$$

$$= \frac{1}{1 - q_i^{t_i}} \sum_{t_{-i}\in T_{-i}\setminus E'_{-i}:h^k_{-i}(t_{-i})=h^k_{-i}(\hat{t}_{-i})} \beta_i(t_{-i}|t_i, h^k_{-i}(t_{-i}) = h^k_{-i}(\hat{t}_{-i})) = 1.$$

Using this notation, the above equation is now

$$\sum_{\hat{t}_{-i}\in\hat{T}_{-i}} \hat{\beta}_i(\hat{t}_{-i}|\hat{t}_i)u_i(m_i', \phi_{-i}(\hat{t}_{-i})), \theta_i, \theta_{-i})$$

Thus, if $\sigma_i(\hat{t}_i)(m_i) > 0$, then, by definition of **p**-dominant equilibrium, it also maximizes this expression. □

1.6.4 p-dominant Implementation and Confident Implementation

In the previous section, we show that in any countable type space, there is an equilibrium in which each agent plays the same action (or sends the same message) as in the **p**-dominant equilibrium in the maximal belief closed subset whenever their type profile is included in the event where the designer's information about agents n-th order belief is **p**-evident. In this subsection, we show such event happens with probability arbitrarily close to 1 conditional on each k-th order belief as ϵ is sufficiently small and when **p** is not too big.

We proceed this in two steps. Recall that under the condition of the distinguishability by k-th order belief, there is a unique $k+1$-th order belief (and higher up to n) of agent i based on the designer' information about agent i. We first show that for any small $\epsilon \geq 0$, for any ϵ-elaboration, the distribution over k-th order belief is "close" to some type space that corresponds to some belief closed subset. That is, we show that the distribution of k-th order belief is close to *some* convex combination of the priors for the minimal consistent belief closed subset. Recall that any types of agent i in the belief closed subset under the condition of distinguishability by k-th order belief has different k-th order belief (see Lemma 1.1). Let $(\hat{T}^\alpha)_{\alpha=1}^{n_\alpha}$ be the consistent minimal belief closed subsets in $E^{*\Delta^n}$ where $n_\alpha \in \mathbb{N}$ is the number of the minimal belief closed subsets (note that it is finite by Lemma 1.1) . For each $\hat{T}^\alpha$, let $\hat{\mathcal{T}}^\alpha = ((\hat{T}_i)^\alpha, \hat{\beta}^\alpha)$ be the common prior type space that corresponds to $\hat{T}^\alpha$.

Lemma 1.3. *Fix $\eta > 0$. Then, there exists $\bar{\epsilon}_\eta > 0$ s.t. for all $\epsilon \leq \bar{\epsilon}_\eta$, for any ϵ-elaboration of $E^{*\Delta^n}$, $\mathcal{T}^\epsilon = ((T_i)_i^\epsilon, \beta^\epsilon)$, we can find $(w_\alpha)_{\alpha=1}^{n_\alpha}$ (it may depend on ϵ and $\mathcal{T}^\epsilon$), where $w_\alpha \in [0,1]$ and $\sum_\alpha w_\alpha = 1$, s.t. for any $\hat{t} \in \hat{T}^\alpha$,*

$$|\beta^\epsilon(\{t \in T^\epsilon : h^k(t) = h^k(\hat{t})\}) - w_\alpha\hat{\beta}^\alpha(\hat{t})| < \eta.$$

Proof. See Appendix. □

Corollary 1.2. *Fix $\eta > 0$. Then, there exists $\bar{\epsilon} > 0$ s.t. for all $\epsilon \leq \bar{\epsilon}$,*

$$\beta^{\epsilon}(t \in T^{\epsilon} : \exists \hat{t} \in \hat{T} \text{ s.t. } h^k(t) = h^k(\hat{t})) > 1 - \eta.$$

Proof. It is proven during the proof of Theorem 1.1, especially in deriving (1.10). □

Note that this does not mean that the distribution of each belief hierarchy converges to that of $\hat{t}$.

In addition, since each k-th order belief has a unique $(k+1)$-th order belief in the designer's information, this means that in an ϵ-elaboration, conditional on each k-th order belief, only n-th order belief that coincides with the designer's information occurs with a high probability.

We exploit the following important result comes from Kajii and Morris (1997) (Proposition 4.2). We slightly extend theirs in a way to allow conditioning on an event. This result provides a connection between the ex-ante probability of an event and the ex-ante probability that the event is common **p**-belief (conditional on an event): if **p** satisfies some condition (roughly, it cannot be too large), then the probability of any event in any type space to be common **p**-belief is arbitrarily close to 1, if the ex-ante probability of that event is arbitrarily close to 1. This result comes at surprising especially taking account Rubinstein (1989).[23]

Given a type space $\mathcal{T} = ((T_i)_i, \beta)$, an event $E \subseteq T$ is *simple* if $E = \prod_{i \in \mathcal{I}} E_i$ where $E_i \subseteq T_i$ for each $i \in \mathcal{I}$. Also for an event F with $\beta(F) > 0$, let $\beta(E|F) := \frac{\beta(E \cap F)}{\beta(F)}$.

Lemma 1.4 (Critical Lemma). *For $\mathbf{p} \in [0,1]^N$, suppose that $\sum_{i \in \mathcal{I}} p_i < 1$ and let $\chi(\mathbf{p}) = \frac{1 - \min_{i \in \mathcal{I}} p_i}{1 - \sum_{i \in \mathcal{I}} p_i}$. Then, for any common prior type space $((T_i)_i, \beta)$, any simple event E and F with $\beta(F) > 0$,*

$$\beta(C^{\mathbf{p}}(E|F)|F) \geq 1 - \chi(\mathbf{p})(1 - \beta(E|F)).$$

The proof is a straightforward extension of Kajii and Morris (1997), applying the same line of arguments except conditioning on F. Thus, we omit it here.

[23] Here the restriction of elaboration to common prior type spaces is crucial.

Theorem 1.1. *Let $E^{*\Delta^n} \subseteq T^*$ be an n-th order belief event for some $n \geq 1$ and distinguishable by k-th order belief for some $k < n$. Then, if a social choice correspondence F is $\mathbf{p}$-dominant implementable where $\sum_{i \in \mathcal{I}} p_i < 1$ in $\hat{T}$ where $\hat{T}$ is the maximal consistent belief closed subset in $E^{*\Delta^n}$, then it is confidently implementable w.r.t. $E^{*\Delta^n}$.*

The crux of the proof is to apply the critical lemma for each event $E^k(\hat{t}) \equiv \{t \in T : h^k(t) = h^k(\hat{t})\}$ "whenever it is possible." The complication arises because the definition of ϵ-elaboration does not rule out the possibility that for some $\hat{t} \in \hat{T}$, the probability $E^k(\hat{t})$ is arbitrarily close to 0, thus the conditional probability of $E^{k+1}(\hat{t})$ conditional on $E^k(\hat{t})$ does not need to be close to 1, which makes applying the critical lemma difficult. The proof circumvents this difficulty in the following way. We divide $\hat{T}$ into two groups, where for any $\hat{t}$ in the first group happens with probability at least some threshold real value (which is chosen to be substantially small); while for $\hat{t}$ in the second group does not. If ϵ is sufficiently small, the conditional probability of $E^{k+1}(\hat{t})$ conditional on $E^k(\hat{t})$ should be close to 1 so for these $\hat{t}$ we apply the critical lemma and our previous result. For the probability of the second group can be made arbitrarily small by choosing sufficiently small threshold. Thus, in total, the probability of achieving the social choice correspondence is arbitrarily close to 1 as ϵ is sufficiently small.

Proof. Fix $\delta > 0$. Let $M \equiv |\hat{T}|$ which is finite by Lemma 1.1. Let $\bar{x}_\delta \equiv \frac{\delta}{4M}$ and $\eta_\delta \equiv \frac{\delta}{4M}$. Note that for any ϵ-elaboration $\mathcal{T} = ((T_i)_i, \beta)$, for any $\hat{t} \in \hat{T}$ s.t. $\beta(E^k(\hat{t})) \geq \bar{x}_\delta$ where $E^k(\hat{t}) \equiv \{t \in T : h^k(t) = h^k(\hat{t})\}$, by definition of ϵ-elaboration,

$$\epsilon \geq \beta(T \setminus E^{k+1}(\hat{t}) | E^k(\hat{t}))\beta(E^k(\hat{t})) \geq \beta(T \setminus E^{k+1}(\hat{t}) | E^k(\hat{t}))\bar{x}_\delta$$

Thus,

$$\beta(T \setminus E^{k+1}(\hat{t}) | E^k(\hat{t})) \leq \frac{\epsilon}{\bar{x}_\delta} \tag{1.8}$$

where $\beta(E'|E) := \frac{\beta(E\cap E')}{\beta(E)}$ for any $E' \subseteq T$ and E with $\beta(E) > 0$. Observer that

$$\begin{aligned}
\sum_{t\in T}\beta(\{t\in T : g(\sigma(t)) \in F(\tilde{\theta}(t))) &\geq \sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta} \beta(\{t\in T : g(\sigma(t)) \in F(\tilde{\theta}(t))|E^k(\hat{t}))\beta(E^k(\hat{t})) \\
&\geq \sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta} \beta(C^{\mathbf{p}}(E^{k+1}(\hat{t})|E^k(\hat{t}))|E^k(\hat{t}))\beta(E^k(\hat{t})) \\
&\geq \sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta} \big(1-\chi(\mathbf{p})(1-\beta(E^{k+1}(\hat{t})|E^k(\hat{t})))\big)\,\beta(E^k(\hat{t})) \\
&\geq \sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta} \left(1-\chi(\mathbf{p})\frac{\epsilon}{\bar{x}_\delta}\right)\beta(E^k(\hat{t})).
\end{aligned}$$

where the second and third inequlities come from Proposition 1.4 and Lemma 1.4, respectively; while the last comes from (1.8). We shall show that this is greater than $1-\delta$ if ϵ is sufficiently small. Note that

$$\sum_{\hat{t}:\beta(E^k(\hat{t}))\leq\bar{x}_\delta} \beta(E(\hat{t})) \leq \bar{x}_\delta|\hat{T}| \leq \frac{\delta}{4M}M = \frac{\delta}{4}. \tag{1.9}$$

In addition, observe that by Lemma 1.3, there exists $\bar{\epsilon}_1 > 0$ s.t. for any $\epsilon \leq \bar{\epsilon}_1$ and each ϵ-elaboration, there exists $(w_\alpha)_{\alpha=1}^{n_\alpha}$ s.t.

$$\begin{aligned}
\left|\sum_{\alpha\in\{1,\ldots,n_\alpha\}}\sum_{\hat{t}\in\hat{T}^\alpha}\Big(\beta(E^k(\hat{t})) - w_\alpha\hat{\beta}^\alpha(\hat{t})\Big)\right| &= \left|\sum_{\alpha\in\{1,\ldots,n_\alpha\}}\sum_{\hat{t}\in\hat{T}^\alpha}\beta(E^k(\hat{t})) - 1\right| \\
&= \left|\sum_{\hat{t}\in\hat{T}}\beta(E^k(\hat{t})) - 1\right| \\
&\leq \sum_{\alpha\in\{1,\ldots,n_\alpha\}}\sum_{\hat{t}\in\hat{T}^\alpha}\left|\beta(E^j(\hat{t})) - w_\alpha\hat{\beta}^\alpha(\hat{t})\right| \\
&\leq |\hat{T}|\eta_\delta = M\frac{\delta}{4M} = \frac{\delta}{4}.
\end{aligned}$$

Thus,

$$\sum_{\hat{t}\in\hat{T}}\beta(E^k(\hat{t})) \geq 1-\frac{\delta}{4}. \tag{1.10}$$

Combining (1.9) and (1.10), we can conclude that if $\epsilon \leq \bar{\epsilon}_1$,

$$\sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta}\beta(E^k(\hat{t})) \geq 1-\frac{\delta}{2}$$

Thus, then we can find $\bar{\epsilon} < \bar{\epsilon}_1$ such that for any $\epsilon \leq \bar{\epsilon}$,

$$\sum_{\hat{t}\in\hat{T}:\beta(E^k(\hat{t}))\geq\bar{x}_\delta} \left(1 - \chi(\mathbf{p})\frac{\epsilon}{\bar{x}_\delta}\right)\beta(E^k(\hat{t})) \geq 1-\delta$$

and as a result,

$$\sum_{t\in T}\beta(\{t\in T : g(\sigma(t)) \in F(\tilde{\theta}(t))) \geq 1-\delta.$$

□

1.7 Discussion

1.7.1 Another Sufficient Condition for Confident Implementability

Definition 1.15 (Liu (2015))**.** Given a common prior type space $\mathcal{T} = ((T_i)_i, \beta)$ and a mechanism $\mathcal{M} = ((M_i)_i, g)$, a recommendation policy $\nu \in \Delta(T \times M)$ is a *belief-invariant Bayes correlated equilibrium (BCE) of* $(\mathcal{M}, \mathcal{T})$ if for each $i \in \mathcal{I}$,

$$\sum_{t\in T}\beta(t)\sum_{m\in M}\nu(m|t)u_i(g(m_i,m_{-i}),\tilde{\theta}(t)) \geq \sum_{t\in T}\beta(t)\sum_{m\in M}\nu(m|t)u_i(g(\phi_i(t_i,m_i),m_{-i}),\tilde{\theta}(t)) \tag{1.11}$$

for all $\phi_i : T_i \times M_i \to M_i$ and for each m_i; and

$$\sum_{m_{-i}}\nu(m_i, m_{-i}|t_{-i}, t_i) \tag{1.12}$$

is independent of t_{-i}.

Proposition 1.5. *Let $\hat{\mathcal{T}} = (\hat{T}, \beta)$ be the maximal consistent type space in E^*. Suppose that a mechanism $\mathcal{M} = (M, g)$ implements a social choice correspondence F in BNE in the maximal consistent belief closed subset in E^*. Suppose also that there exists a unique belief-invariant BCE of $(\mathcal{M}, \hat{\mathcal{T}})$. Then, F is confidently implementable w.r.t. E^*.*

It is known that belief-invariant BCE is invariant (Theorem 2 of Liu (2015)), i.e., it only depends on belief hierarchy on Θ.

Lemma 1.5. *For any $\eta > 0$, there exists $\bar{\epsilon} > 0$ such that for any ϵ-elaboration $\mathcal{T} = (T, \beta)$ with $\epsilon \in [0, \bar{\epsilon}]$ and any equilibrium $\sigma \equiv (\sigma_i)_i$ of $(\mathcal{M}, \mathcal{T})$,*

$$\nu(m,\hat{t}) := \sum_{t \in T: h(t)=\hat{t}} \beta(t) \prod_{i \in I} \sigma_i(m_i|t_i), \forall m \in M, \forall \hat{t} \in \hat{T}.$$

is a η-belief-invariant BCE of $(\mathcal{M}, \hat{\mathcal{T}})$.

$$\sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu(m|t^*) u_i(g(m), \tilde{\theta}(t^*)) - \sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu(m, t^*) u_i(g(\phi_i(t_i^*, m_i), m_{-i}), \tilde{\theta}(t^*)) \geq -\eta \tag{1.13}$$

and satisfies (1.12).

Proof. Fix $\eta > 0$ and let $\epsilon > 0$ be small enough so that $\epsilon < \frac{\eta}{2B}$ where B is the bound for utility i.e., $|u_i(x,\theta)| \leq B$ for all i, x and θ. Consider an ϵ-elaboration, $\mathcal{T} = (T, \beta)$

$$\sum_{t \in T} \beta(t) u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})) \geq \sum_{t \in T} \beta(t) u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$

Let $E \equiv \{t \in T : h(t) \in E^*\}$. Then,

$$\sum_{t \in E} \beta(t) u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})) + \sum_{t \in T \setminus E} \beta(t) u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$

$$\geq \sum_{t \in E} \beta(t) u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})) + \sum_{t \in T \setminus E} \beta(t) u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$

Note that

$$\begin{aligned}\sum_{t \in T \setminus E} \beta(t) \left(u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}(t)) - u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}(t)) \right) &\leq \sum_{t \in T \setminus E} \beta(t) 2B \\ &\leq \epsilon 2B < \eta.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\sum_{t \in E} \beta(t) u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})) \\ &\qquad - \sum_{t \in E} \beta(t) u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})) \geq -\eta \end{aligned} \tag{1.14}$$

Note that without loss $t_i = (t_i^*, s_i)$ where $t_i^* = h_i(t_i)$ and $s_i \in S_i$ for some S_i. That is, we can decompose it into the payoff relevant private information and irrelevant one. Let $S \equiv \prod_{i \in \mathcal{I}} S_i$.

Define $\nu : T^* \to \Delta(M)$

$$\nu(m|t^*) = \frac{1}{\hat{\beta}(t^*)} \sum_{s \in S} \beta(t^*, s) \prod_{i \in I} \sigma_i(m_i|t_i), \forall m \in M, \forall \hat{t} \in \hat{T} \tag{1.15}$$

where

$$\hat{\beta}(t^*) \equiv \sum_{s \in S} \beta(t^*, s).$$

Note that given t^*, $\sum_{m \in M} \nu(m|t^*) = 1$.

First we show that the obedience condition (1.11) holds. Suppose not. Then, there exists $\phi_i : T_i \times M_i \to M_i$ s.t.

$$\begin{aligned}
&\sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu(m, t|t^*) u_i(g(m_i, m_{-i}), \tilde{\theta}(t^*)) + \eta < \sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu(m, t|t^*) u_i(g(\phi_i(t_i, m_i), m_{-i}), \tilde{\theta}(t^*)) \\
&\iff \sum_t \sum_m \beta(t) \sigma(m|t) u_i(g(m_i, m_{-i}), \tilde{\theta}(t)) + \eta < \sum_t \sum_m \beta(t) \sigma(m|t) u_i(g(\phi_i(t_i, m_i), m_{-i}), \tilde{\theta}(t)) \\
&\iff \sum_t \sum_m \beta(t) u_i(g(\sigma_i(t_i), \sigma_{-i}(t_{-i}), \tilde{\theta}(t)) + \eta < \sum_t \sum_m \beta(t) u_i(g(\sigma_i'(t_i), \sigma_{-i}(t_{-i}), \tilde{\theta}(t))
\end{aligned}$$

where $\sigma_i'(m_i'|t_i) := \sum_{m_i \in M_i} \sigma_i(m_i|t_i) \phi_i(m_i'|t_i, m_i)$; this is a contradiction to (1.14). Thus we have

$$\sum_{t^* \in E^*} \frac{\hat{\beta}(t^*)}{\hat{\beta}(E^*)} \sum_{m \in M} \nu(m|t^*) u_i(g_i(m), \tilde{\theta}(t^*)) \geq \sum_{t^* \in E^*} \frac{\hat{\beta}(t^*)}{\hat{\beta}(E^*)} \sum_{m \in M} \nu(m|t^*) u_i(g(\phi_i(t_i^*, m_i), m_{-i}), \tilde{\theta}(t^*))$$

Let

$$\tilde{T}_i := \{t_i^* \in T_i^* : \exists t_{-i}^* \in T_{-i}^* \text{ s.t. } \hat{\beta}(t_i^*, t_{-i}^*) > 0\}$$

Then, $((\tilde{T}_i)_i, \frac{\hat{\beta}}{\hat{\beta}(E)})$ be a common prior type space in $\hat{T}$. We also need to check (1.12).

$$\begin{aligned}
\sum_{m_{-i} \in M_{-i}} \nu(m_i, m_{-i}|t^*) &= \frac{1}{\hat{\beta}(t_i^*, t_{-i}^*)} \sum_{s_i \in S_i, s_{-i} \in S_{-i}} \beta((t_i^*, s_i), (t_{-i}^*, s_{-i})) \sigma_i(m_i|t_i^*, s_i) \\
&= \frac{1}{\hat{\beta}(t_i^*, t_{-i}^*)} \sum_{s_i \in S_i} \left(\sum_{s_{-i} \in S_{-i}} \beta(t_{-i}^*, s_{-i}|t_i^*, s_i) \right) \Pr(t_i^*, s_i) \sigma_i(m_i|t_i^*, s_i)
\end{aligned}$$

where

$$\Pr(t_i^*, s_i) \equiv \sum_{t_{-i}^*, s_{-i}} \beta((t_i^*, s_i), (t_{-i}^*, s_{-i})) \text{ and } \beta(t_{-i}^*, s_{-i}|t_i^*, s_i) \equiv \frac{\beta((t_i^*, s_i), (t_{-i}^*, s_{-i}))}{\Pr(t_i^*, s_i)}.$$

Since

$$\sum_{s_{-i} \in S_{-i}} \beta(t_{-i}^*, s_{-i}|t_i^*, s_i) = \hat{\beta}(t_{-i}^*|t_i^*)$$

by the definition of t^* and s,

$$\begin{aligned}\sum_{m_{-i} \in M_{-i}} \nu(m_i, m_{-i}|t_i^*, t_{-i}^*) &= \frac{1}{\hat{\beta}(t_i^*, t_{-i}^*)} \hat{\beta}(t_{-i}^*|t_i^*) \sum_{s_i \in S_i} \Pr(t_i^*, s_i)\sigma_i(m_i|t_i^*, s_i) \\ &= \frac{1}{\Pr(t_i^*)} \sum_{s_i \in S_i} \Pr(t_i^*, s_i)\sigma_i(m_i|t_i^*, s_i)\end{aligned}$$

which is independent of t_{-i}^* as desired.

□

Corollary 1.3. *Assume E^* is countable. Consider a sequence $(\epsilon^k)_k$ such that $\epsilon^k \to 0$ and ϵ^k-elaboration of E^*, $\mathcal{T}^k = ((T_i^k)_i, \beta^k)$. For each k consider a BNE $\sigma^k \equiv (\sigma_i^k)_i$ of $(\mathcal{M}, \mathcal{T}^k)$. Then, $(\sigma^k)_k$ converges to a belief-invariant BCE of $(\mathcal{M}, \hat{\mathcal{T}})$.*

Proof. Let $(\eta_l)_l$ with $\eta_l \to 0$. By Lemma 1.5, we can find a subseqeunce k_l such that for each l

$$\begin{aligned}&\sum_{t^* \in E^*} \hat{\beta}^{k_l}(t^*) \sum_{m \in M} \nu^{k_l}(m|t^*) u_i(g(m), \tilde{\theta}(t^*)) \\ &\qquad\qquad \geq \sum_{t^* \in E^*} \hat{\beta}^{k_l} \sum_{m \in M} \nu^{k_l}(m, t^*) u_i(g(\phi_i(t_i^*, m_i), m_{-i}), \tilde{\theta}(t^*)) - \eta_l\end{aligned}$$

As we assume that E^* is countable and $\hat{\beta}^{k_l}(t^*)$ and $\nu^{k_l}(m|t^*)$ is clearly bounded, there exists a subsequence such that

$$\hat{\beta}^{k_l}(t^*) \to \hat{\beta}(t^*), \forall t^* \in E^*$$

$$\nu^{k_l}(m|t^*) \to \nu(m|t^*), \forall m \in M, t^* \in E^*$$

and

$$\sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu^{k_l}(m|t^*) u_i(g(m), \tilde{\theta}(t^*)) \geq \sum_{t^* \in E^*} \hat{\beta}(t^*) \sum_{m \in M} \nu(m, t^*) u_i(g(\phi_i(t_i^*, m_i), m_{-i}), \tilde{\theta}(t^*)).$$

□

Now prove the proposition.

Proof. Suppose not. Then, we can find some $\delta > 0$ and a sequence $(\epsilon^k)_k$ where $\epsilon^k \to 0$ such that there exists some BNE σ^k of $(\mathcal{M}, \mathcal{T}^k)$ such that

$$\beta^k(t \in T^k : g(\sigma^k(t)) \notin F(\theta)) > \delta$$

for each k. By define $\nu^k : T^* \to \Delta(M)$ as in (1.15),

$$\sum_{t^* \in E*} \hat{\beta}^k(t^*) \sum_{m \in M: g(m) \notin F(\tilde{\theta}(t^*))} \nu^k(m|t^*) + \sum_{t^* \in T \setminus E*} \hat{\beta}^k(t^*) \sum_{m \in M: g(m) \notin F(\tilde{\theta}(t^*))} \nu^k(m|t^*) > \delta$$

For sufficiently large k

$$\frac{1}{\hat{\beta}(E^*)} \sum_{t^* \in E*} \hat{\beta}^k(t^*) \sum_{m \in M: g(m) \notin F(\tilde{\theta}(t^*))} \nu^k(m|t^*) > \delta$$

Then, by Corollary 1.3, there exists a convergent subsequence

$$\sum_{t^* \in E^*} \hat{\beta}(t) \sum_{m \in M: g(m) \notin F(\tilde{\theta}(t^*))} \nu(m|t^*) \geq \delta \tag{1.16}$$

Note that the unique belief-invariant in the maximal belief closed subset should be a BNE $\sigma^* \equiv (\sigma_i^*)_i$, because we assume countable type space, there exists always a BNE. Since $\mathcal{M}$ implements F in the maximal consistent belief closed subset, this unique BNE should satisfy

$$\hat{\beta}(t \in \hat{T} : g(\sigma^*) \in F(\tilde{\theta}(t))) = 1$$

In other words,

$$\sum_{t^* \in T^*} \hat{\beta}(t^*) \sum_{m \in F(\tilde{\theta}(t^*)} \prod_{i \in I} \sigma_i(m_i|t_i^*) = \sum_{t^* \in T^*} \hat{\beta}(t^*) \sum_{m \in M: g(m) \in F(\tilde{\theta}(t^*)} \nu(m|t^*) = 1,$$

which is a contradiction to (1.16). □

1.8 Conclusion

In this paper, we introduce a novel notion of robustness into mechanism design theory which we call confident implementation, and provide a framework to study this. We also introduce **p**-dominant implementation and show that when the designer has information about agents' n-th order belief, **p**-dominant implementability with certain range of **p** is a sufficient condition for a social choice correspondence to be confidently implementable with respect to the designer's information. Also, using this characterization, we identify social choice correspondences that are confidently implementable but not ex-post implementable.

We conclude this paper by providing future directions which we are pursuing. In this paper, we only consider common prior type spaces for ϵ-elaboration. We are working on the cases of allowing non-common prior type spaces. Also, we are working for more results for **p**-dominant implementability, especially to know to which sense or to the extent of **p**-dominant implementability is also necessary for confident implementability or the other robustness foundations we suggested.

1.9 Appendix

1.9.1 Omitted Proofs

1.9.1.1 Proof of Proposition 1.1

Proof. Fix $\delta > 0$. Since F is confidently implementable w.r.t. $E^{*'}$, there exists $\bar{\epsilon} > 0$ such that for any $\epsilon \leq \bar{\epsilon}$ and any $\mathcal{T} \in \mathcal{E}(E^{*'}, \epsilon)$ theres exists σ such that

$$\beta(\{t \in T : g(\sigma(t)) \in F(\tilde{\theta}(t))\}) \geq 1 - \delta.$$

Observe that since $E^* \subseteq E^{*'}$, for any $\mathcal{T}$, $\beta(E) \leq \beta(E')$. Thus, for any ϵ, if $\mathcal{T} \in \mathcal{E}(E^*, \epsilon)$, then $\mathcal{T} \in \mathcal{E}(E^{*'}, \epsilon)$. Let $\epsilon \leq \bar{\epsilon}$ and consider $\mathcal{T} \in \mathcal{E}(E, \epsilon)$. Then, by the observation, $\mathcal{T} \in \mathcal{E}(E^{*'}, \epsilon)$, and we know there exists σ that satisfies the above condition. □

1.9.1.2 Proof of Lemma 1.3

Proof. Suppose not. Then we can find some sequence $(\epsilon^l)_{l=0}^{\infty}$ such that $\epsilon^l \to 0$ and some sequence of ϵ^l-elaboration of $E^{*\Delta^n}$, $\mathcal{T}^l = ((T_i^l)_i, \beta^l)$ s.t. for any $(w_\alpha)_{\alpha=1}^{n_\alpha}$ with $w_\alpha \geq 0$ and

$\sum_{\alpha=1}^{n_\alpha} w_\alpha = 1$,

$$|\beta^l(\{t \in T^l : h^k(t) = h^k(\hat{t})\}) - w_\alpha \hat{\beta}^\alpha(\hat{t})| \geq \eta. \tag{1.17}$$

Consider a subsequence of $(\mathcal{T}^l)_{l=1}^\infty$ such that

$$\beta^l(\{t \in T^l : h^k(t) = h^k(\hat{t})\}) \to \beta'(\hat{t}) \in [0,1], \forall \hat{t} \in \hat{T}.$$

(we use the same notation for the subsequence for notational convenience). Such a subsequence exists because 1) $\beta^l(\{t \in T^l : h^k(t) = h^k(\hat{t})\}) \in [0,1]$ for each l and $[0,1]$ is compact; and 2) $\hat{T}$ is finite. Moreover, for each ϵ^l-elaboration,

$$1 \geq \beta^l(\{t \in T : h(t) \in E^{*\Delta^n}\}) = \beta^l(\{t \in T : \exists \hat{t} \in \hat{T} \text{ s.t. } h^k(t) = h^k(\hat{t})\}) \geq 1 - \epsilon^l$$

where the equality comes from the regularity of $E^{*\Delta^n}$ (Assumption 1.1) and Lemma 1.1. Hence, $\beta' \in \Delta(\hat{T})$.

Claim 1.1. Given $\alpha \in \{1, \ldots, n_\alpha\}$, suppose that $\hat{t}_i \in \hat{T}_i^\alpha$ satisfies

$$\sum_{\hat{t}_{-i} \in \hat{T}_{-i}} \beta'(\hat{t}_i, \hat{t}_{-i}) > 0 \tag{1.18}$$

and

$$h_i^{k+1}(\hat{t}_i) = (\theta_i, \delta_i^{k+1}) \in T_i^{*k+1}.$$

Then,

$$\beta_i'(\hat{t}_{-i}|\hat{t}_i) := \frac{\beta'(\hat{t}_i, \hat{t}_{-i})}{\sum_{\hat{t}_{-i} \in \hat{T}_{-i}} \beta'(\hat{t}_i, \hat{t}_{-i})} = \hat{\beta}_i^\alpha(\hat{t}_{-i}|\hat{t}_i) = \delta_i^{k+1}(\hat{t}_{-i}), \forall \hat{t}_{-i} \in \hat{T}_{-i}. \tag{1.19}$$

In words, the conditional probability of the limit distribution, $\hat{\beta}_i'(\cdot|\hat{t}_i)$ coincides with $\beta_i^\alpha(\cdot|\hat{t}_i)$ if $\hat{t}_i \in \hat{T}_i^\alpha$.

Proof. Given ϵ^l-elaboration, $\mathcal{T}^l = ((T_i)_i, \beta^l)$, denote

$$E_i^k(\hat{t}_i) \equiv \{t_i \in T_i : h_i^k(t_i) = h^k(\hat{t}_i)\}, \forall i \in \mathcal{I}, \forall k, \forall \hat{t}_i \in \hat{T}_i.$$

To see (1.19),

$$\begin{aligned}
\beta_i'(\hat{t}_{-i}|\hat{t}_i) &\equiv \frac{\beta^l(E_i^k(\hat{t}_i)\times E_{-i}^k(\hat{t}_{-i}))}{\beta^l(E_i^k(\hat{t}_i)\times T_{-i}^l)}+\epsilon_0^l \\
&= \frac{\beta^l(E_i^{k+1}(\hat{t}_i)\times E_{-i}^k(\hat{t}_{-i}))+\beta^l((E_i^k(\hat{t}_i)\setminus E_i^{k+1}(\hat{t}_i))\times E_{-i}^k(\hat{t}_{-i}))}{\beta^l(E_i^{k+1}(\hat{t}_i)\times T_{-i}^l)+\beta^l((E_i^k(\hat{t}_i)\setminus E_i^{k+1}(\hat{t}_i))\times T_{-i}^l)}+\epsilon_0^l \\
&= \frac{\beta^l(E_i^{k+1}(\hat{t}_i)\times E_{-i}^k(\hat{t}_{-i}))+\epsilon_1^l}{\beta^l(E_i^{k+1}(\hat{t}_i)\times T_{-i})+\epsilon_2^l}+\epsilon_0^l \\
&= \delta_i^{k+1}(\hat{t}_{-i})+\epsilon_3^l.
\end{aligned}$$

for some small positive ϵ_0^l, ϵ_1^l, ϵ_2^l and ϵ_3^l. The third equality holds because $\mathcal{T}^l$ is ϵ^l-elaboration:

$$\begin{aligned}
\beta^l((E_i^k(\hat{t}_i)\setminus E_i^{k+1}(\hat{t}_i))\times E_{-i}^k(\hat{t}_{-i})) &\leq \beta^l((E_i^k(\hat{t}_i)\setminus E_i^{k+1}(\hat{t}_i))\times T_{-i}^l) \\
&\leq \beta^l(\{t\in T : h(t)\notin E^{*\Delta^n}\}) \\
&\leq \epsilon^l,
\end{aligned}$$

where the second inequality comes from the assumption the distinguishability with k-th order belief.

Let us prove the last equality. Observe that for any $t_i \in E_i^{k+1}(\hat{t}_i)$

$$\beta_i^l(E_{-i}^{\hat{t}_{-i}}|t_i) \equiv \frac{\beta^l(\{t_i\}\times E_{-i}^{\hat{t}_{-i}})}{\beta^l(\{t_i\}\times T_{-i})} = \delta_i^{k+1}(\hat{t}_{-i}).$$

Equivalently,

$$\beta^l(\{t_i\}\times E_{-i}^k(\hat{t}_{-i})) = \delta_i^{k+1}(\hat{t}_{-i})\beta^l(\{t_i\}\times T_{-i}), \forall t_i \in E_i^{k+1}(\hat{t}_i).$$

From this

$$\sum_{t_i\in E_i^{k+1}(\hat{t}_i)} \beta^l(\{t_i\}\times E_{-i}^k(\hat{t}_{-i})) = \delta_i^{k+1}(\hat{t}_{-i}) \sum_{t_i\in E_i^{k+1}(\hat{t}_i)} \beta^l(\{t_i\}\times T_{-i}), \forall t_i \in E_i^{k+1}(\hat{t}_i)$$

Simply

$$\beta^l(E_i^{k+1}(\hat{t}_i)\times E_{-i}^k(\hat{t}_{-i})) = \delta_i^{k+1}(\hat{t}_{-i})\beta^l(E_i^{k+1}(\hat{t}_i)\times T_{-i}).$$

Hence,

$$\frac{\beta^l(E_i^{k+1}(\hat{t}_i)\times E_{-i}^k(\hat{t}_{-i}))}{\beta^l(E_i^{k+1}(\hat{t}_i)\times T_{-i})} = \delta_i^{k+1}(\hat{t}_{-i}).$$

As we assume (1.18), the denominator $\beta^l(E_i^k(\hat{t}_i)\times T_{-i})+\epsilon_2^l > 0$ for sufficiently large l. Thus, “0/0 situation” does not happen. □

Claim 1.2. Given some $\alpha \in \{1, \ldots, n_\alpha\}$, suppose there exists $\hat{t} \in \hat{T}^\alpha$ s.t.

$$\beta'(\hat{t}) > 0.$$

Then, for any $\hat{t}' \in \hat{T}^\alpha$,

$$\beta^\alpha(\hat{t}') > 0 \iff \beta'(\hat{t}') > 0.$$

Proof. Since $\beta'(\hat{t}) > 0$. For each $i \in \mathcal{I}$, by Claim 1.1,

$$\beta'_i(\cdot|\hat{t}_i) = \hat{\beta}_i^\alpha(\cdot|\hat{t}_i).$$

In particular, it should be true that

$$\beta'_i(\hat{t}'_{-i}|\hat{t}_i) > 0 \iff \hat{\beta}_i^\alpha(\hat{t}'_{-i}|\hat{t}_i) > 0.$$

Hence,

$$\beta'(\hat{t}_i, \hat{t}'_{-i}) > 0 \iff \hat{\beta}^\alpha(\hat{t}_i, \hat{t}'_{-i}) > 0$$

for any $(\hat{t}_i, \hat{t}'_{-i})$. If there is no $\hat{t}'_{-i} \neq \hat{t}_{-i} \in \hat{T}_{-i}^\alpha$ such that $(\hat{t}_i, \hat{t}'_{-i}) \in \operatorname{supp} \hat{\beta}^\alpha$, we stop. Otherwise, we apply the same argument for some $(\hat{t}_i, \hat{t}'_{-i}) \in \operatorname{supp} \hat{\beta}^\alpha$. And so on. □

For each $\alpha \in \{1, \ldots, n_\alpha\}$, define

$$w_\alpha := \begin{cases} \sum_{\hat{t} \in \hat{T}^\alpha} \beta'(\hat{t}) & \text{if } \exists \hat{t} \in \hat{T}^\alpha \text{ s.t. } \beta'(\hat{t}) > 0 \\ 0 & \text{o.w.} \end{cases}.$$

Claim 1.3. Given some $\alpha \in \{1, \ldots, n_\alpha\}$, suppose there exists $\hat{t} \in \hat{T}^\alpha$ s.t.

$$\beta'(\hat{t}) > 0.$$

Then, $\frac{1}{w_\alpha}\beta' \in \Delta(\hat{T}^\alpha)$ and

$$\frac{1}{w_\alpha}\beta'(\hat{t}) = \hat{\beta}^\alpha(\hat{t}), \forall \hat{t} \in \hat{T}^\alpha.$$

Proof. Recall that for each α, $\hat{\beta}^\alpha$ is a common prior for a minimal consistent belief closed subset. It is known that such a common prior is unique (Corollary 4.7 of Mertens and Zamir (1985)). By Claim 1.1 and by Claim 1.2, we know that $\frac{1}{w_\alpha}\beta'$ is a common prior for $(\hat{\beta}_i^\alpha)_{i \in \mathcal{I}}$ for the same minimal consistent belief closed subset; hence, it should be the same as $\hat{\beta}^\alpha$. □

Note that

$$\sum_{\alpha=1}^{n_\alpha} w_\alpha = 1.$$

By the series of claims above we show that

$$\beta^l(\{t :\in T : h^k(t) = h^k(\hat{t})\}) \to w_\alpha \beta^\alpha(\hat{t}).$$

This is a contradiction to (1.17). □

1.9.1.3 Proof of Proposition 1.2

Proof. Consider agent the incentive compatibility of agent i with type t_i. By definition of **p**-dominant equilibrium,

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(\sigma_i(t_i), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$

$$\geq \sum_{t_{-i}} \beta_i(t_{-i}|t_i) u_i(g(m_i', \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall m_i' \in M_i$$

for any $\phi_{-i}(t_{-i}) \in \Delta(M_{-i})$ such that each player $j \neq i$ with t_j plays $\sigma_j(t_j)$ with probability at least p_i and arbitrarily with the rest probability. In particular,

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(\sigma_i(t_i), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$

$$\geq \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(\sigma_i(t_i'), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i' \in T_i$$

This implies

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(\sigma_i(t_i), \phi_{-i}'(t_{-i})), \tilde{\theta}_i t_i, \tilde{\theta}_{-i}(t_{-i}))$$

$$\geq \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(\sigma_i(t_i'), \phi_{-i}'(t_{-i})), \tilde{\theta}(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i \in T_i'$$

where $\phi_{-i}'(t_{-i}) \in \Delta(M_{-i})$ where each $j \neq i$ plays $\sigma_j(t_j)$ with at least probability p_i and plays arbitrarily m_j from

$$M_j' \equiv \{m_j \in M_j : \exists t_j \in T_j, \sigma_j(t_j)[m_j] > 0\}.$$

That is, the set of messages which are sent by some t_j' in σ_j with positive probability.

Define

$$f(t_i, t_{-i}) := g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \forall t_i \in T_i, t_{-i} \in T_{-i}.$$

Then the above inequality implies

$$\sum_{t_{-i} \in T_{-i}} \beta_i(t_{-i}|t_i) u_i(f(t_i, \tilde{\phi}_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$
$$\geq \sum_{t_{-i} \in T_{-i}} \beta_i(t_{-i}|t_i) u_i(f(t_i', \tilde{\phi}_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i' \in T_i \quad (1.20)$$

for any $\tilde{\phi}_{-i}(t_{-i})$ such that any $j \neq i$ reports truthfully at least probability p_i. Lastly, observe that

$$f(t) = g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

□

CHAPTER 2

p-dominant Implementation

2.1 Introduction

Strategic uncertainty (that is, uncertainty about others' actions and others' belief and higher order beliefs about others' action) of agents has been regarded as an important consideration in designing robust mechanisms.[1] In the private value case in which each agent's valuations to alternatives are fully determined by his/her own private information about payoff-relevant parameters, requiring dominant strategy equilibrium is most robust solution concept one can employ in this regard in that each agent's equilibrium strategy is a best response *regardless of* the other agents' strategy. By doing so, such mechanism is robust even when some agents may not be fully rational or make mistakes.

On the other hand, in the interdependent value case, where each agent's value to items may depend on the other agents' private information, ex-post equilibrium (Crémer and McLean, 1985) has been regarded as a corresponding notion to dominant strategy equilibrium. This notion requires that incentive compatibility holds for any realization of the other agents' types. Especially, when we consider direct mechanisms, in the private value case, the two notions are equivalent since in this case for each agent's action space is his/her type space. In this sense, ex-post equilibrium is a generalization of dominant strategy equilibrium to the interdependent value case.

However, ex-post equilibrium does not capture strategic uncertainty which dominant strategy equilibrium does. This is because, ex-post equilibrium requires that each agent's

[1] Refer to Yamashita (2015) and references therein for more about strategic uncertainty and structural uncertainty.

equilibrium strategy is a best response *only* to the other agents' equilibrium strategy. Instead, this notion captures robustness to uncertainty in information structures of agents by requiring the incentive compatibility to hold for any realization of (payoff) types. In the literature, such an uncertainty is called structural uncertainty.

The main contribution of the present paper is to provide a notion in the interdependent value case which captures robustness to strategic uncertainty only. We provide such a notion by extending **p**-dominant equilibrium (Morris et al., 1995; Kajii and Morris, 1997) to incomplete information games.[2] Importantly, this notion is defined with respect to a given type space in contrast to ex-post equilibrium; so this notion does not concern informational robustness. This notion is defined as follows: a strategy profile is **p**-dominant equilibrium in a type space where $\mathbf{p} \in [0,1]^N$ when there are $N \in \mathbb{N}$ agents, if for each agent, playing equilibrium strategy is a best response as long as the other agents play their equilibrium strategy with probability at least p_i (and arbitrarily play for the rest probability).

We provide a logical relationship between this notion and the existing well known concepts. As can be easily seen, when $\mathbf{p} = \mathbf{1}$, the notion reduces to Bayes Nash equilibrium. When $\mathbf{p} = \mathbf{0}$, this notion is different from both dominant strategy equilibrium and ex-post equilibrium. More precisely, for general $\mathbf{p} \neq \mathbf{1}$, in particular $\mathbf{p} = 0$, ex-post equilibrium and **p**-dominant equilibrium are logically orthogonal. In the literature, in the interdependent value case, dominant strategy equilibrium is defined to capture both strategic uncertainty and informational uncertainty by requiring incentive compatibility holds regardless of the other agents' play *and* types.[3] Thus, in the interdependent value case, dominant strategy equilibrium is stronger than the other two notions.

Our second main contribution is to completely characterize **p**-dominant implementable allocation in the quasilinear environment with a single item to be allocated in which each agent's private information, called payoff-type, is one-dimensional and each agent's valuation is a weighted sum of their own payoff type and the others'. To be more precise, we assume

[2] See also relevant concepts p-best response (Tercieux, 2006); and p-rationalizability (Hu, 2007).

[3] See, for example, Definition 5 of Bergemann and Morris (2005).

common prior payoff type spaces and independence of types. In this environment, it is well-known that Bayesian implementable allocation is fully characterized by the monotonicity condition, which requires that each agent's interim expected allocation rule is (weakly) increasing in each agent's reported type. To obtain an extension of this result,[4] we first define $\mathbf{p}$-monotonicity, which requires each agent i's interim expected allocation to be increasing in his report for any reporting strategy of the opponents for which they truthfully report with probability at least p_i.

We find that $\mathbf{p}$-monotonicity alone does not characterize $\mathbf{p}$-dominant implementation allocation. More precisely, the former is necessary but not sufficient for the latter; and we find that we need an additional condition for this. This additional condition requires that the independence of agent's valuation is sufficiently small.

We then turn to $\mathbf{p}$-dominant implementable allocation in continuous type spaces. Perhaps surprisingly, we find that for any $\mathbf{p} \in [0,1]^N$ for which $p_i < 1$ for all i, $\mathbf{p}$-dominant incentive compatibility is equivalent to $\mathbf{0}$-dominant incentive compatibility. Thus, we only need to focus on $\mathbf{0}$-dominant implementable allocation rules in this case. Given this observation, we make a similar characterization to the one in the discrete type case, but a stronger form: even with arbitrarily small interdependence, $\mathbf{0}$-dominant implementable allocation requires the derivative of the allocation with respect to each agent's report to be independent of the other agents' report. Using this characterization, we also characterize the constrained efficient $\mathbf{p}$-dominant implementable allocation when there are two agents.

The concept of $\mathbf{p}$-dominant implementation naturally captures strategic uncertainty of agents. The degree of strategic uncertainty of agent i is represented by p_i. We provide formal foundations for $\mathbf{p}$-dominant implementation: in Chapter 1 we already provide one: a social choice correspondence is confidently implementable with respect to some subset E^* of the universal type space if there exists a mechanism such that for any type spaces in which the higher order belief of a realized type profile is sufficiently likely to be in E^*, there exists a Bayes Nash equilibrium whose outcome is an element of the social choice correspondence

[4] Recall that when $\mathbf{p} = 1$, $\mathbf{p}$-dominant equilibrium coincides with BNE.

with probability arbitrarily close to 1. We then show that if a social choice correspondence is $\mathbf{p}$-dominant implementable in the maximal belief closed subset in E^* where $\sum_i p_i < 1$ is confident implementable with respect to E^*. We also provide two more formal foundations for $\mathbf{p}$-dominant implementation.

Crémer and McLean (1985) studies ex-post implementable allocation rules in the quasi-linear environment. They find that an allocation rule is ex-post implementable if and only if it satisfies ex-post monotonicity. Comparing to this, $\mathbf{p}$-dominant implementable allocation is completely characterized by $\mathbf{p}$-monotonicity and the extra condition that restricts the interdependence. Since $\mathbf{0}$-monotonicity coincides with ex-post monotonicity, ex-post monotonicity is most stringent in terms of the monotonicity, i.e., ex-post monotonicity implies any $\mathbf{p}$-monotonicity. However, due to the presence of the extra condition on the interdependence, in general, ex-post implementable allocation does not need to be $\mathbf{p}$-dominant implementable, and vice versa.

Some similar conditions on the interdependence can be found in the literature. For example, in Bergemann and Morris (2009a), for an efficient allocation to be robust implementable in their sense, the interdependence should not be too large (see also Ollár and Penta (2017)). Interestingly, they study full implementation and the condition bites for this. For partial implementation version of their robustness concept in Bergemann and Morris (2005), such a condition does not appear, as long as ex-post monotonicity holds. On the other hand, in this paper we studies partial implementation; but still it requires the interdependence condition.

The rest of the paper is organized as follows. In Section 2.2, we present the model. In Section 2.3, we introduce the main concept of this paper, $\mathbf{p}$-dominant implementation. We provide some preliminary result and introduce $\mathbf{p}$-monotonicity in Section 2.4. We completely characterize $\mathbf{p}$-dominant implementable allocations in Section 2.5 and Section 2.6 in discrete type spaces and continuous type spaces, respectively. We provide a couple of robustness foundation in Section 2.7; then, conclude the paper in Section 2.8 with a brief discussion of future directions.

2.2 Setting

2.2.1 Environment

There is a mechanism designer ("she") and there is a finite set of agents $\mathcal{I} = 1, 2, \ldots, N$ (each of them is called "he"). There is a finite set of alternatives X. Each agent's preference over $\mathcal{X} \equiv \Delta(X)$ satisfies conditions for the expected utility representation and depends on payoff-relevant parameters. Each agent has private information about the parameters which is called *payoff type* of agent i and let Θ_i be the set of payoff types for agent i with a typical element of θ_i.

The designer's goal is represented by a social choice correspondence $F : \Theta \rightrightarrows \mathcal{X}$.

In later sections, we shall focus on the *quasilinear environment* with a single item to be assigned (e.g., auction): $\mathcal{X} = [0,1]^N \times \mathbb{R}^N$ with a typical element of (q, τ); here, $q : \Theta \to [0,1]^N$ and $\tau : \Theta \to \mathbb{R}^N$ are called *allocation rule* and *transfer rule*, respectively. An allocation rule q is feasible if $\sum_{i \in \mathcal{I}} q_i(\theta) \leq 1$ for any $\theta \in \Theta$. In addition, each agent's utility is given as

$$u_i((q,\tau),\theta) = v_i(\theta)q_i + \tau_i, \forall (q,\tau), \forall \theta \in \Theta$$

where $v_i : \Theta \to \mathbb{R}$ which call *valuation* of agent i.[5]

In this environment, we assume that the designer only cares about implementable allocation rules (see Definition 2.8).

2.2.2 Type Space

A *type space* is a tuple $((T_i)_{i\in\mathcal{I}}, (\tilde{\beta}_i)_{i\in\mathcal{I}}, (\tilde{\theta}_i)_{i\in\mathcal{I}})$ where $\tilde{\beta}_i : T_i \to \Delta(T_{-i})$ and $\tilde{\theta}_i : T_i \to \Theta_i$ for each i. A *payoff type space* is a type space where $\tilde{\theta}_i$ is a bijection for every i.

Note that in a type space, there may be two types whose payoff types are the same but have different beliefs, i.e., there exist $t_i, t_i' \in T_i$ and $\tilde{\beta}_i(t_i) = \tilde{\beta}_i'(t_i')$ while $\tilde{\theta}_i(t_i) \neq \tilde{\theta}_i(t_i')$.

[5] Note that each agent is risk neutral in money. This justifies our restriction of $\mathcal{X}$ to $[0,1]^N \times \mathbb{R}^N$.

2.2.3 Mechanism and Implementation

Definition 2.1. A *mechanism (a game form)* is a pair $((M_i)_{i \in \mathcal{I}}, g)$ where M_i is a nonempty set for each $i \in \mathcal{I}$ and $g : M \to \mathcal{X}$.

We call M_i the *message space* for agent i and call g the *outcome function*. Note that (M, g) may be an extensive-form.

A particularly simple class of mechanisms is *direct* mechanisms. In a direct mechanism agents are supposed to report their type, i.e., $M_i = T_i$ for each $i \in \mathcal{I}$.

A type space $\mathcal{T}$ and a mechanism $\mathcal{M}$ induce a Bayesian game $(\mathcal{M}, \mathcal{T})$.

Given a Bayesian game $(\mathcal{M}, \mathcal{T})$, a strategy of agent i, σ_i, is defined as a mapping from T_i to $\Delta(M_i)$.

Definition 2.2. Given $(\mathcal{M}, \mathcal{T})$, a strategy profile $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ is a *Bayes Nash equilibrium* (henceforth BNE) if for each $i \in \mathcal{I}$, $t_i \in T_i$ and $m_i \in M_i$ with $\sigma_i(t_i)[m_i] > 0$,

$$m_i \in \arg\max_{m_i' \in M_i} \sum_{t_{-i} \in T_{-i}} \beta_i(t_{-i}|t_i) u_i(g(m_i, \sigma_{-i}(t_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})).$$

Definition 2.3. A mechanism (M, g) *(partially) implements in BNE* a social choice correspondence F in a common prior type space $\mathcal{T}$, if there exists a Bayes Nash equilibrium $\sigma = (\sigma_i)_i$ such that for any $t \in T$ s.t. $\beta(t) > 0$,

$$g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

And we call such F is *(partially) implementable in BNE*. In words, the notion requires the existence of an equilibrium that yields the desirable outcome for each realization of payoff type profile.[6]

[6] The notion of partial implementation is different from *full implementation*, which requires *every* equilibrium to achieve the desirable outcome.

2.3 p-dominant Implementation

In this section, we first introduce **p**-dominant equilibrium in Bayesian games and the corresponding implementation notion. Then we provide a revelation principle for this new notion of implementation.

2.3.1 p-dominant Equilibrium in Bayesian Games

We extend **p**-dominant equilibrium (Morris et al., 1995; Kajii and Morris, 1997), which was originally defined in complete information games, to games with incomplete information in order to employ it in mechanism design.[7] There may be potentially more ways to extend the notion; the reason why we chose this way will be clear shortly.

Definition 2.4. Let $\mathbf{p} \in [0,1]^N$. Given a game $(\mathcal{T}, \mathcal{M})$, a strategy profile $\sigma \equiv (\sigma_i)_i$ where $\sigma_i : T_i \to \Delta(M_i)$ is a **p**-*dominant equilibrium* if for each $i \in \mathcal{I}, t_i \in T_i$ and $m_i \in M_i$ with $\sigma_i(t_i)[m_i] > 0$,

$$m_i \in \arg\max_{m_i' \in M_i} \sum_{t_{-i}} \beta_i(t_{-i}|t_i) u_i(g(m_i', \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})))$$

for any $\phi_{-i} : T_{-i} \to \Delta(M_{-i})$ such that for each $t_{-i} \in T_{-i}$

$$\phi_{-i}(t_{-i}) = q_i^{t_{-i}} \sigma_{-i}(t_{-i}) + (1 - q_i^{t_{-i}}) \psi_{-i}(t_{-i}) \tag{2.1}$$

for some $q_i^{t_{-i}} \geq p_i$ and $\psi_{-i}(t_{-i}) \in \Delta(M_{-i})$.

In words, a strategy profile constitutes **p**-dominant equilibrium if for each agent i and t_i, the equilibrium strategy is a best response to any conjecture over the opponents' message profiles that puts on probability at least p_i on the equilibrium strategy profile; for the rest probability $1 - p_i$, the opponents' strategies are allowed to be correlated across agents (but not correlated within types of an agent). We will call ψ_{-i} in (2.1) *babbling* of $-i$.

[7] As noted in Morris et al. (1995), the notion of p-dominance is a generalization of Harsayni and Selten's risk-dominance in 2×2 games in the sense that it coincides risk dominance when $\mathbf{p} = (1/2, 1/2)$.

In addition, given a mechanism, if σ is a $\mathbf{p}$-dominant equilibrium; then it is also a $\mathbf{p}'$-dominant equilibrium for any $\mathbf{p}' \geq \mathbf{p}$.[8] In particular, any $\mathbf{p}$-dominant equilibrium is a Bayesian Nash equilibrium. Clearly, $\mathbf{p}$-dominant equilibrium may not exist.[9] With private value, when $\mathbf{p} = \mathbf{0}$, this notion is equivalent to (weakly) dominant strategy equilibrium. However, it shall be shown momentarily in Section 2.3.3 that the notion is weaker than dominant strategy equilibrium with interdependent value. In particular, we should emphasize that the set of $\mathbf{p}$-dominant equilibrium depends on the underlying type space $\mathcal{T}$, even when $\mathbf{p} = 0$.

Definition 2.5. A social choice correspondence $F : \Theta \rightrightarrows \mathcal{X}$ is $\mathbf{p}$-*dominant implementable* in a type space $\mathcal{T}$ if there exists a mechanism $\mathcal{M} = (M, g)$ and a $\mathbf{p}$-dominant equilibrium σ in $(\mathcal{M}, \mathcal{T})$ such that for each $t \in T$

$$g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

Note that it is a refinement of partial implementation in BNE (Definition 2.3), simply because a $\mathbf{p}$-dominant equilibrium is a Bayes Nash equilibrium.

2.3.2 Revelation Principle for p-dominant Implementation

There are infinite number of mechanisms to be checked in order to see whether a social choice correspondence is $\mathbf{p}$-dominant implementable. In the (partial) implementation in BNE, the revelation principle allows us to focus on the direct mechanism for this purpose. In this subsection, we make a parallel observation for $\mathbf{p}$-dominant implementability.

Proposition 2.1 (Revelation principle for $\mathbf{p}$-dominant implementation)**.** *Let $\mathcal{M} = (M, g)$ be a mechanism, and let $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ where $\sigma_i : T_i \to \Delta(M_i)$ be a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}, \mathcal{T})$. Then there exists a direct mechanism $\mathcal{M}' = ((T)_{i \in \mathcal{I}}, f)$ such that*

(1) Truthful reporting, i.e., $\sigma_i'(t_i) = t_i$ for all $i \in \mathcal{I}$, is a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}', \mathcal{T})$.

[8] $\mathbf{p}' \geq \mathbf{p}$ if each $p_i' \geq p_i$ for all $i \in \mathcal{I}$.

[9] In this regard, see also relevant concepts (p-BR, p-MBR) in Tercieux (2006).

(2) For every $t \in T$,

$$f(t) = g(\sigma(t)).$$

Note that if $g(\sigma(t)) \in F(\tilde{\theta}(t))$, then $f(t) \in F(\tilde{\theta}(t))$. The argument is quite standard; if anything is nonstandard, it would be the part dealing with the babbling of the opponents. Note that with the original indirect mechanism, agents have more messages to send in the sense that in the induced direct mechanism the messages corresponding to an agent's reported types is a subset of the message space of the original indirect mechanism, so agent i effectively needs to consider a smaller set of babbling of the opponents in the indirect mechanism.

Proof. See Appendix. □

Due to this result, from now on we focus on direct mechanisms when we consider **p**-dominant implementability.

2.3.3 Discussion: Dominant Strategy Equilibrium, Ex-post Equilibrium and 0-dominant Equilibrium

As previously mentioned, we extend the existing notion of **p**-dominant equilibrium, which is initially defined in complete information games, to games with incomplete information games. In doing so, we find that **0**-dominant equilibrium is logically orthogonal to ex-post equilibrium; and also different from dominant strategy equilibrium generally. In this subsection, we discuss their relationships. For this purpose, let us first define the two other solution concepts formally.

Definition 2.6. A direct mechanism $((\Theta_i)_{i\in\mathcal{I}}, f)$ is *ex-post incentive compatible* if for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$,

$$u_i(f(\theta_i, \theta_{-i}), \theta_i, \theta_{-i}) \geq u_i(f(\theta_i', \theta_{-i}), \theta_i, \theta_{-i}), \forall \theta_i' \in \Theta_i, \theta_{-i} \in \Theta_{-i}.$$

Definition 2.7. A direct mechanism $((\Theta_i)_{i\in\mathcal{I}}, f)$ is *dominant strategy incentive compatible* if for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$,

$$u_i(f(\theta_i, \theta_{-i}'), \theta_i, \theta_{-i}) \geq u_i(f(\theta_i', \theta_{-i}'), \theta_i, \theta_{-i}), \forall \theta_i' \in \Theta_i, \theta_{-i} \in \Theta_{-i}, \theta_{-i}' \in \Theta_{-i}.$$

F	θ_h	θ_l
θ_h	1	0
θ_l	0	0

Figure 2.1: Social choice function in Example 2.1

By the revelation principles for each equilibrium concept, a social choice correspondence is implementable in each solution concept if there exists a direct mechanism that achieves it.[10]

In the following example, we show that there is a social choice correspondence that is ex-post implementable but not **0**-dominant implementable.

Example 2.1. There are two agents $i \in \{1, 2\}$ and for each i, let $\Theta_i = \{\theta_h, \theta_l\}$ where $\theta_h = 1$ and $\theta_l = -2$. Assume that types are independently drawn across agents and let $\Pr(\theta_h) = \lambda \in [0, 1]$. Consider a social choice correspondence (in fact function) in Figure 2.1.

Let $X \equiv \{0, 1\}$ where $x = 1$ represents build a public good; while $x = 0$ represents not building it. We assume that agents have interdependent value with $v_i(\theta_i, \theta_j) = \theta_i + \theta_j$ for each i, $j \neq i$.

Let us first see that the direct mechanism (Θ, f) where $f = F$ is ex-post implementable: to see this, consider incentive compatibility of agent i with θ_h:

$$u_i(f(\theta_h, \theta_h), \theta_h, \theta_h) = 1 > u_i(f(\theta_l, \theta_h), \theta_h, \theta_h) = 0$$

$$u_i(f(\theta_h, \theta_l), \theta_h, \theta_l) = 0 = u_i(f(\theta_l, \theta_l), \theta_h, \theta_l) = 0$$

For agent i with θ_l, it is weakly dominant to truthfully report. Thus, this social choice function is ex-post implementable.

To show that this social choice function is not **0**-dominant implementable it is sufficient to show that there exists a babbling $\psi_j : \Theta_j \to \Theta_j$, $j \neq i$, with which incentive compatibility

[10] The employed definitions come from Bergemann and Morris (2005) (see also their footnote 10 for more references for ex-post equilibrium). We define relevant implementability roughly. Refer to the same paper for the exact definitions of them.

of agent i is violated. Consider the following babbling: $\psi_j(\theta_h) = \theta_l$ and $\psi_j(\theta_l) = \theta_h$, i.e., each type of agent j reports the opposite type. In this case, when agent i reports truthfully his utility is

$$\lambda u_i(f(\theta_h, \theta_l), \theta_h, \theta_h) + (1 - \lambda) u_i(f(\theta_h, \theta_h), \theta_h, \theta_l)) < 0$$

Thus, truth-telling is not **0**-dominant equilibrium; thus the social choice function is not **0**-dominant implementable (note that we use the revelation principle for **p**-dominant implementation Proposition 2.1 here).

Proposition 2.2. *Let $((\Theta_i)_{i\in\mathcal{I}}, f)$ be a direct mechanism. There are following logical relationship between dominant strategy equilibrium, ex-post equilibrium and* **0***-dominant equilibrium.*

(1) With private value, three solution concepts are equivalent.

(2) With interdependent value, any dominant strategy equilibrium strategy profile is an ex-post equilibrium. But, the converse is not true.

(3) With interdependent value, any dominant strategy equilibrium strategy profile is a **0**-dominant equilibrium. But, the converse is not true.

(4) With interdependent value, neither ex-post equilibrium implies nor implied by **0**-dominant equilibrium.

(5) There is a game and a strategy profile which is both ex-post and **0**-dominant equilibrium but not a dominant strategy equilibrium.

See Figure 2.2 for a schematic exposition of the proposition.

Proof. For (1), the equivalence of ex-post equilibrium and dominant strategy equilibrium with private value is pointed out by Bergemann and Morris (2005). Thus we need to only show the equivalence of dominant strategy equilibrium and **0**-dominant equilibrium with private value. This is simply because the additional incentive constraints by varying the opponents' payoff type do not bite by definition of private value.

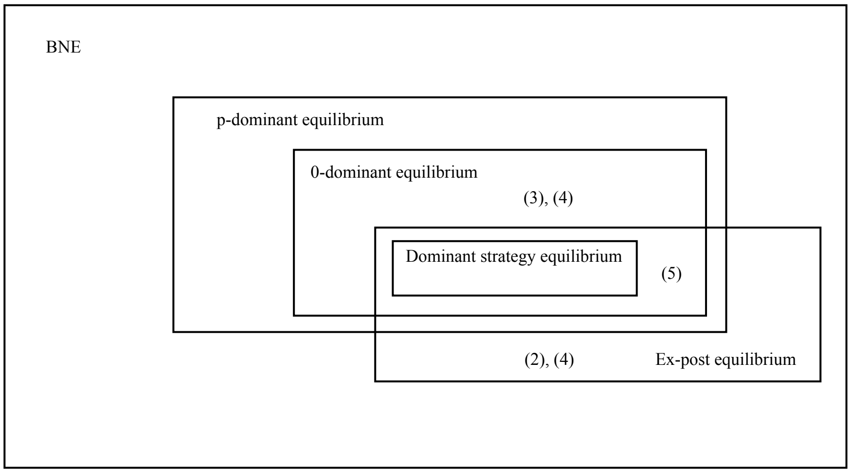


Figure 2.2: A schematic exposition of Proposition 2.2

It is clear that dominant strategy equilibrium is stronger than ex-post equilibrium and **0**-dominant equilibrium. For the second part of (2), as dominant strategy equilibrium is stronger than **0**-dominant equilibrium, it is enough to show that there is an ex-post equilibrium but not **0**-dominant equilibrium, which is already shown in Example 2.1.

For the second part of (3), as dominant strategy equilibrium is stronger than ex-post equilibrium, it is sufficient to show there is **0**-dominant equilibrium that is not ex-post equilibrium. We already have seen such case through Example 1.1 when $\lambda \geq 1/3$ in the example.

We have already shown (4) in proving (2) and (3).

We prove (5) in Example 2.2. □

Example 2.2. Suppose that there are two agents $i = 1, 2$. Their payoff type is given by $\Theta_i = \{\theta_l, \theta_h\}$ where $\theta_l = 0$ and $\theta_h = 1/2$. Agents' type is independently and identically drawn; denote $\lambda \equiv \Pr(\theta_h)$. Let $X = \{0, 1/2, 1\}$ and for each agent i,

$$u_i(x, (\theta_i, \theta_j)) = -|x - (\theta_i + \theta_j)|, \forall x \in X, \theta_i \in \Theta_i, \theta_j \in \Theta_j$$

and the social choice function is given by Figure 2.3.

Truthful-reporting of their type is an ex-post equilibrium, because the social choice function assigns the best alternative for each payoff type profile.

F	θ_h	θ_l
θ_h	1	1/2
θ_l	1/2	0

Figure 2.3: Social choice function in Example 2.2

We next claim that truthful-reporting is **0**-dominant equilibrium if and only if $\lambda = 1/2$. To see this, consider incentive of player 1 with type θ_h. Consider agent 2's babbling $\psi_2(\theta_h) = \theta_l$ and $\psi_2(\theta_l) = \theta_h$. In this case, the expected utility is

$$\begin{aligned} -\lambda\left|\frac{1}{2} - (\theta_h + \theta_h)\right| - (1-\lambda)|1 - (\theta_h + \theta_l)| &= -\frac{1}{2} \\ &\geq -\lambda|0 - (\theta_h + \theta_h)| - (1-\lambda)\left|\frac{1}{2} - (\theta_h + \theta_l)\right| = -\lambda. \end{aligned}$$

Now consider the incentive compatibility of agent 1 with type θ_l with the same babbling of agent 2:

$$\begin{aligned} -\lambda|0 - (\theta_l + \theta_h)| - (1-\lambda)\left|\frac{1}{2} - (\theta_l + \theta_l)\right| &= -\frac{1}{2} \\ &\geq -\lambda\left|\frac{1}{2} - (\theta_l + \theta_h)\right| - (1-\lambda)|1 - (\theta_l + \theta_l)| = -(1-\lambda). \end{aligned}$$

Thus both incentive compatibilities hold only when $\lambda = 1/2$. We can check incentive compatibilities for the other babbling of agent 2 in a similar way and can find when $\lambda = 1/2$ they hold.

As **p**-dominant equilibrium is defined with respect to a given type space, when $\lambda = 1/2$, truth-telling is both ex-post and **0**-dominant equilibrium. However, it is not dominant strategy equilibrium, because it does not satisfy incentive compatibility for agent 1 with θ_h, for example, when agent 2's type is θ_l and reports θ_h

$$-|1 - (\theta_h + \theta_l)| = -\frac{1}{2} < -\left|\frac{1}{2} - (\theta_h + \theta_l)\right| = 0.$$

2.4 Preliminary Result and p-monotonicity in Quasilinear Environment

2.4.1 Basic Characterization

The following simple observation significantly reduces the number of incentive compatibility conditions to be checked.

Lemma 2.1. *$f : T \to \mathcal{X}$ satisfies* $\mathbf{p}$*-dominant incentive compatibility (i.e., (2.12)) if and only if for each i, $t_i \in T_i$ and $\psi_{-i} : T_{-i} \to T_{-i}$.*

$$t_i \in \arg\max_{t_i' \in T_i} p_i \left(\sum_{t_{-i}} \beta_i(t_{-i}|t_i) u_i(f(t_i', t_{-i}), (t_i, t_{-i})) \right) + (1 - p_i) \left(\sum_{t_{-i}} \beta_i(t_{-i}|t_i) u_i(f(t_i', \psi_{-i}(t_{-i})), t_i, t_{-i}) \right).$$

Proof. See Appendix. □

Recall that the original definition of $\mathbf{p}$-dominant incentive compatibility requires that truthful-reporting is a best response to any conjecture about the opponents report with at least p_i of truthful reporting. This proposition reduces the number of incentive compatibilities in two ways: first, it is only needed to check when the opponents *exactly* reports truthfully p_i and with the rest probability they arbitrarily report; second, for the arbitrary report part, it is sufficient to consider "pure" report of each type.

Due to this characterization, henceforth we will only consider the simplified incentive compatibility.

We characterize the set of allocation rules that are $\mathbf{p}$-dominant implementable in the quasilinear environment. We study both private value and interdependent value cases.

We first study here discrete type case. Especially, there is a single item to be allocated to N agents: $\mathcal{X} = [0, 1]^N$ and $\Theta_i = \{\theta_i^0, \theta_i^1, \ldots, \theta_i^{K_i}\} \subseteq \mathbb{R}$ where $\theta_i^0 \leq \theta_i^1 \cdots \leq \theta_i^{K_i}$ for each $i \in \mathcal{I}$ where $K_i \in \mathbb{N} \cup \{0\}$.

Throughout this section, we maintain the following assumptions:

Assumption 2.1 (One-dimensional payoff type space)**.** *For each agent i, $\Theta_i \subseteq \mathbb{R}$. In addition, $T_i = \Theta_i$ and $\tilde{\beta}_i(\cdot|\theta_i) = \lambda_i(\cdot|\theta_i) \in \Delta(\Theta_{-i})$ and $\tilde{\theta}_i(\theta_i) = \theta_i, \forall i \in \mathcal{I}, \theta_i \in \Theta_i$.*

Assumption 2.2 (Linearly interdependent utility)**.** *Each agent i's valuation $v_i : \Theta \to \mathbb{R}$ has the following form:*

$$v_i(\theta) = \theta_i + \gamma \sum_{j \neq i} \theta_j, \forall \theta \in \Theta$$

where $\gamma \geq 0$.

Namely, each agent's valuation for the item is a weighted sum of the agent's own payoff type and the sum of the other agents' payoff types.

Definition 2.8. An allocation rule $q = (q_i)_{i \in \mathcal{I}}$ is **p**-*dominant implementable* if for each agent i, there exists $\tau_i : \Theta \to \mathbb{R}$ which makes (q, τ) satisfies **p**-dominant incentive compatibility.

2.4.2 p-monotonicity

Given an allocation rule $(q_i)_i$ where $q_i : \Theta \to [0,1]$, $\mathbf{p} \in [0,1]^N$ and a function $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$, denote

$$Q_i^{p_i}(\theta_i, \psi_{-i}) := p_i \left(\sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) q_i(\theta_i, \theta_{-i}) \right) \\ + (1 - p_i) \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) q_i(\theta_i, \psi_{-i}(\theta_{-i})), \forall \theta_i \in \Theta_i.$$

That is, $Q_i^{p_i}(\theta_i, \psi_{-i})$ is the interim expected allocation of agent i with type θ_i when the opponents truthfully report with probability p_i and babble according to ψ_{-i} with the rest probability.

We introduce the following concept:

Definition 2.9. An allocation rule $q = (q_i)_{i \in \mathcal{I}}$ satisfies **p**-*monotonicity* if for each i, $Q_i^{p_i}(\theta_i, \psi_{-i})$ is increasing in θ_i for any $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$.

In words, this definition requires that for any (pure) babbling of the opponents, the expected allocation is increasing in agent i's report. Note that when $p_i = 1$ for every $i \in \mathcal{I}$, this definition reduces to the standard monotonicity for implementability in BNE.

By the following observation, it is sufficient to check the monotonicity of babbling "uniformly" for checking the **p**-monotonicity. We define *uniform babbling* of agent j if agent j reports the same report regardless of his true type, i.e.,

$$\psi_j(\theta'_j) = \theta_j, \forall \theta'_j \in \Theta_j.$$

In this case, we denote

$$Q_i^{p_i}(\theta_i, \theta_{-i}) \equiv Q_i^{p_i}(\theta_i, \psi_{-i})$$

where $\psi_{-i}(\theta'_{-i}) = \theta_{-i}$ for all θ'_{-i}.

Lemma 2.2. *An allocation rule $q = (q_i)_{i \in \mathcal{I}}$ satisfies* **p***-monotonicity if and only if for each agent i, $Q_i^{p_i}(\theta_i, \theta_{-i})$ is increasing in θ_i for any $\theta_{-i} \in \Theta_{-i}$.*

Proof. See Appendix. □

The following is immediate from the lemma.

Corollary 2.1. **0***-monotonicity is equivalent to ex-post monotonicity.*

2.5 Characterization of p-dominant Implementability with Discrete Payoff type Spaces

Definition 2.10. A payoff type space $((\Theta_i)_{i \in \mathcal{I}}, \lambda)$ is *independent* if for each $\theta \in \Theta$,

$$\lambda(\theta) = \prod_{i \in \mathcal{I}} \lambda_i(\theta_i), \forall \theta \in \Theta$$

where $\lambda_i(\theta_i) \equiv \sum_{\theta_{-i}} \lambda_i(\theta_i, \theta_{-i})$.

We assume that $((\Theta_i)_{i \in \mathcal{I}}, \lambda)$ is independent for this subsection.

2.5.1 Private Value

It turns out that the analysis for the interdependent value case is substantially more complicated. Thus, we first study the private value case where each agent's valuation of the item does not depend on others' type. Formally, agent i's valuation is *private* if

$$v_i(\theta_i, \theta_{-i}) = \theta_i, \forall \theta_{-i} \in \Theta_{-i}.$$

Proposition 2.3. *Suppose that each agent's valuation is private. Then, an allocation rule* $q \equiv (q_i)_{i \in \mathcal{I}}$ *is* $\mathbf{p}$*-dominant implementable if and only if* q *satisfies* $\mathbf{p}$*-monotonicity.*

Proof. This result is subsumed by Theorem 2.1; hence the proof is omitted. □

In words, the only condition that bites for an allocation rule to be $\mathbf{p}$-dominant implementable is $\mathbf{p}$-monotonicity in the private value case as it is the case in implementability in BNE. Note that this includes the standard characterization of Bayesian implementability as a special case $\mathbf{p} = 1$.

2.5.2 Interdependent Value

Lemma 2.3. *In this environment, an allocation rule* $q = (q_i)_{i \in \mathcal{I}}$ *is* $\mathbf{p}$*-dominant implementable, then* q *satisfies* $\mathbf{p}$*-monotonicity.*

Proof. See Appendix. □

The proof is standard except that we need to consider incentive constraints for *every* babbling of other agents: suppose $\mathbf{p}$-monotonicity does not hold. This means that there is some babbling of the opponents, and the expected allocation given this babbling violates the monotonicity, i.e., there are two types for which the lower type's expected allocation is higher than that for the higher type; then by incentive compatibility, the lower type prefers this higher allocation and payment, which implies the higher types does due to the single crossing property. A contradiction.

Is $\mathbf{p}$-monotonicity also sufficient as it is in the private value case? It turns out that it is not the case. The following example illustrates this.

Example 2.3. In this example, we show that $\mathbf{0}$-monotonicity is not sufficient for an allocation rule to be $\mathbf{p}$-dominant implementable. Let $N = 2$ and $\Theta_1 = \{\theta_h, \theta_l\}$ where $\theta_h > \theta_l$,

$\Theta_2 = \{\theta_u, \theta_d\}$ where $\theta_u > \theta_d$. Types are independently drawn and equally likely. Let $\gamma = 1$, i.e., $v_i = \theta_i + \theta_j$, $j \neq i$. Let us first introduce the following notations:

$$\Delta q_1(\theta_u) \equiv q_1(\theta_h, \theta_u) - q_1(\theta_l, \theta_u)$$

; and similarly for $\Delta q_1(\theta_l)$ Similarly, define

$$\Delta t_1(\theta_u) = t_1(\theta_h, \theta_u) - t_1(\theta_l, \theta_u)$$

The ICs for agent 1 is as follows: when agent 2 uniformly babbles θ_u

$$\left(\theta_h + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \Delta q_1(\theta_u) \geq t_1(\theta_u) \geq \left(\theta_l + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \Delta q_1(\theta_u)$$

and uniformly babbles θ_d

$$\left(\theta_h + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \Delta q_1(\theta_d) \geq \Delta t_1(\theta_d) \geq \left(\theta_l + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \Delta q_1(\theta_d)$$

From these,

$$\begin{aligned}\bar{R} &\equiv \left(\theta_h + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) \\ &\geq \frac{1}{2}\Delta t_1(\theta_u) + \frac{1}{2}\Delta t_1(\theta_d) \\ &\geq \left(\theta_l + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right) \left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) \equiv \bar{L}.\end{aligned}$$

Now consider other types of babbling in which each type reports a different type, i.e., θ_h reports θ_l; and θ_l reports θ_h. In this case,

$$\begin{aligned}R(\psi_2) &\equiv \theta_h \left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) + \frac{1}{2}\theta_u \Delta q_1(\theta_d) + \frac{1}{2}\theta_d \Delta q_1(\theta_u) \\ &\geq \frac{1}{2}\Delta t_1(\theta_u) + \frac{1}{2}\Delta t_1(\theta_d) \\ &\geq \theta_l \left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) + \frac{1}{2}\theta_u \Delta q_1(\theta_d) + \frac{1}{2}\theta_d \Delta q_1(\theta_u) \equiv L(\psi_2)\end{aligned}$$

Note that the middle expression of these inequalities and the above are the same. It should be

$$\max\{L(\psi_2), \bar{L}\} \leq \min\{R(\psi_2), \bar{R}\}$$

Otherwise, the intersection of the above inequalities is empty. Suppose

$$\Delta q_1(\theta_u) > \Delta q_1(\theta_d)$$

and all of them are greater than 0, i.e., it satisfies 0-monotonicity. To see which is smaller between $R(\psi_2)$ and $\bar{R}$, note that for making it smaller, "anti-assortative" is better; i.e., match bigger θ_u with smaller $\Delta q_1(\theta_d)$.

$$\begin{aligned}R(\psi_2) &= \theta_h\left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) + \frac{1}{2}\theta_u\Delta q_1(\theta_d) + \frac{1}{2}\theta_d\Delta q_1(\theta_u) \\ &\geq \bar{L} = \left(\theta_l + \frac{1}{2}\theta_u + \frac{1}{2}\theta_d\right)\left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right)\end{aligned}$$

if and only if

$$(\theta_h - \theta_l)\left(\frac{1}{2}\Delta q_1(\theta_u) + \frac{1}{2}\Delta q_1(\theta_d)\right) \geq \frac{1}{2}\Delta q_1(\theta_u)(\mathbb{E}[\theta_2] - \theta_d) + \frac{1}{2}\Delta q_1(\theta_d)(\mathbb{E}[\theta_2] - \theta_u).$$

Suppose

$$\Delta q_1(\theta_u) = 10 > \Delta q_1(\theta_d) = 1$$

and

$$\theta_u = 10, \theta_d = 0$$

so $\mathbb{E}[\theta_2] = 5$ Then the inequality is

$$(\theta_h - \theta_l)\frac{1}{2}(10 + 1) \geq \frac{1}{2}10 \times 5 + \frac{1}{2}1(-5) = \frac{45}{2}.$$

If $\theta_h - \theta_l$ is small enough this inequality is trivially not satisfied. So, there is no t_1 satisfies all the inequalities above.

Recall that the basic intuition involving monotonicity is that an allocation rule that satisfies the monotonicity provides a larger marginal benefit of reporting a higher type when the true type is higher. This allows us to come up with the transfer rule by which the marginal cost of reporting a higher type to be between those marginal benefits; as a result, providing incentive to report the true type. Here, we need to consider incentive constraints for every babbling of other agents, and the previous example shows that sometimes it is impossible to find a transfer rule that is consistent to every incentive constraint. We precisely characterize how such additional consideration restricts the implementable allocations in the next result.

Theorem 2.1. *An allocation rule $q \equiv (q_i)_{i \in \mathcal{I}}$ is* $\mathbf{p}$*-dominant implementable if and only if*

(1) q satisfies $\mathbf{p}$-monotonicity

(2) For each $i \in \mathcal{I}$ and $\theta_i^k, \theta_i^{k-1} \in \Theta_i$,

$$\theta_i^k - \theta_i^{k-1} \geq \gamma \max_{\psi_{-i}} \left| \sum_{\theta_{-i}} \frac{(1-p_i)\lambda(\theta_{-i})\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))}{p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \sum_{j \neq i} (\theta_j - \mu_j) \right|. \quad (2.2)$$

Proof. See Appendix. □

Comparing to the condition for the standard Bayesian implementability, there is an additional condition, the second condition involving γ and q. It turns out that this condition exactly comes from the fact that we need to consider every babbling possibility with the interdependent value: with the interdependent value, agents' payoff depends on the true types of other agents; even though an agent has the same belief over the other agents' report, it is possible that he has different beliefs over which types are more likely to report a certain type. For instance in Example 2.3, an agent has the belief that the other agent reports each type equally likely. However, he might have belief that each agent reports truthfully report; or exactly opposite report; and this makes his marginal benefit from reporting high type different. Note that for both beliefs, the expected increase in payment is the same. This implies that the difference in the marginal benefit between high and low types should be large enough to take into account all the belief about the other agents' report.

A closer look at the second condition suggests which determines the size of this impact of interacting interdependent value and babbling. Obviously, smaller γ makes it smaller. Especially, when $\gamma = 0$, the condition does not bite; thus, the monotonicity is the only qualification for $\mathbf{p}$-dominant implementability as we already discussed in the previous subsection. Perhaps more interestingly, the second condition involves the marginal increase of an agent's allocation reporting a higher type (i.e., $\Delta q_i(\theta_i, \theta_{-i})$) rather than the absolute level of this. Intuitively, given a babbling of the opponents, if the other agents' report makes big difference in the marginal increase in the allocation, then the impact of the babbling combining with different true types on the difference of the marginal benefit is large.

Corollary 2.2. *For any $i \in \mathcal{I}$ and θ_i^k, if $\Delta q_i(\theta_i^k, \theta_{-i}) = \Delta q_i(\theta_i^k, \theta'_{-i})$ for all $\theta_{-i}, \theta'_{-i}$, then the second condition in the proposition does not bite.*

Proof. See Appendix. □

Given γ, it might not be easy to check whether q satisfies the sufficient condition. The following proposition provides a sufficient condition on γ which can be applied regardless of q. That is, under this sufficient condition, **p**-monotonicity is enough.

Corollary 2.3 (A uniform sufficient condition)**.** *Suppose*

$$\gamma \leq \min_{i \in \mathcal{I}} \min_k (\theta_i^k - \theta_i^{k-1}) \left(\frac{p_i}{1-p_i} \frac{\min_{\theta_{-i}} \lambda_i(\theta_{-i})}{\max_{\theta_{-i}} \lambda_i(\theta_{-i})} + 1 \right) \frac{1}{\max_{\theta_{-i}} \left| \sum_{j \neq i} (\theta_j - \mu_j) \right|}.$$

Then, q is **p***-dominant implementable if and only if q satisfies* **p***-monotonicity.*

Proof. See Appendix. □

The uniform sufficient condition may be unnecessarily strong for a particular q, but it is convenient (as we will see in Example 2.4) .

2.5.2.1 Constrained Efficient p-dominant Allocation Rule

Consider the single item environment. An allocation rule $q^* \equiv (q_i^*)_i$ is *efficient* if

$$q^*(\theta) \in \arg\max \sum_{i \in \mathcal{I}} v_i(\theta) q_i^*(\theta)$$

where the maximization is taken over *feasible* allocations, i.e., $q_i : \Theta \to [0,1]$ for each i and $\sum_i q_i \leq 1$. An allocation rule $q = (q_i)_{i \in \mathcal{I}}$ where $q_i : \Theta \to [0,1]$ is *constant* if for any $i \in \mathcal{I}$ and $\theta \in \Theta$, $q_i(\theta) = k_i$ for some $k_i \in [0,1]$.

Lemma 2.4. *If $\gamma < 1$, then an efficient allocation satisfies ex-post monotonicity. If $\gamma > 1$, then a non-constant efficient allocation violates ex-post monotonicity.*

Proof. See Appendix. □

Intuitively, when $\gamma < 1$, efficiency requires the item to be assigned to an agent with a higher payoff type; while when $\gamma > 1$, giving it to a lower payoff type is more efficient. Since ex-post monotonicity implies $\mathbf{p}$-monotonicity for any $\mathbf{p} \in [0,1]^N$, we have the following result.

Corollary 2.4. *If $\gamma < 1$, an efficient allocation rule satisfies $\mathbf{p}$-monotonicity for any $\mathbf{p} \in [0,1]^N$.*

Lemma 2 of Crémer and McLean (1985) shows that if q satisfies ex-post monotonicity, then it is ex-post implementable. Combining this with Lemma 2.4, we have the following result.

Corollary 2.5. *An (non-constant) efficient allocation rule is ex-post implementable if and only if $\gamma \leq 1$.*

Example 2.4. In this example, we characterize the most efficient allocation rule that is $\mathbf{p}$-dominant implementable for each $\mathbf{p} \in [0,1]$.[11] There are two agents $i \in \{1,2\}$ and for each agent i, $v_i(\theta) = \theta_i + 2\theta_j$, $j \neq i$. Suppose $\theta_i \in \{1, 15\}$ for each i and equally probable.

Note that to maximize the social surplus the item should be assigned to the agent with *lower* type. But this allocation rule violates the Bayesian monotonicity (see Figure 2.4). The constrained efficient allocation rule that is implementable in BNE and ex-post implementable (also implementable in dominant strategy equilibrium) are also described in the same figure.

Given some $\mathbf{p} \in [0,1]^N$, what is the constrained efficient allocation rule that is $\mathbf{p}$-dominant implementable? Due to Theorem 2.1, we can obtain it by solving the following optimization problem:

$$\max_{\{w_i,x_i,y_i,z_i\}_{i=1}^2} \frac{1}{4}((3z_1 + 31y_1 + 17x_1 + 45w_1) + (3z_2 + 17y_2 + 31x_2 + 45w_2))$$

$$= \max_{\{w_i,x_i,y_i,z_i\}_{i=1}^2} \frac{1}{4}\left(3(z_1+z_2) + 31(y_1+x_2) + 17(x_1+y_2) + 45(w_1+w_2)\right)$$

[11] We adapt an example in Gershkov et al. (2013) in a way to allow $\mathbf{p}$-dominant implementable with interior $\mathbf{p}$. They originally characterize most efficient allocation rule for BNE and ex-post equilibrium.

q_1	θ_h	θ_l
θ_h	1/2	0
θ_l	1	1/2

q_1	θ_h	θ_l
θ_h	1/2	1/4
θ_l	3/4	0

q_1	θ_h	θ_l
θ_h	1/2	1/2
θ_l	1/2	1/2

Figure 2.4: The left one is the efficient allocation; the middle one is the most efficient implementable allocation in BNE; the right one is the most efficient ex-post implementable allocation

subject to 1) the well-defined as an allocation rule, i.e.,

$$w_1 + w_2 \leq 1, x_1 + x_2 \leq 1, y_1 + y_2 \leq 1, z_1 + z_2 \leq 1$$

$$w_i, x_i, y_i, z_i \geq 0$$

for $i = 1, 2$ and 2) the **p**-monotonicity (Definition 2.9) and 3) the second condition in Theorem 2.1. For notations, see below.

q_i	θ_h	θ_l
θ_h	w_i	x_i
θ_l	y_i	z_i

To make our exercise simpler let us focus on when $p_1 = p_2 = p$. We first claim that the second condition in Theorem 2.1 is satisfied in this case regardless of q. To see this, we use the uniform version of the sufficient condition (i.e., Corollary 2.3): this amounts to

$$\gamma \leq (\theta_h - \theta_l)\left(\frac{p}{1-p}\frac{\min\{\lambda_h, \lambda_l\}}{\max\{\lambda_h, \lambda_l\}} + 1\right)\frac{1}{\max\{\theta_h - \mu, \mu - \theta_l\}}$$

Since $\lambda_h = \lambda_l = 1/2$, $\mu = \frac{1}{2}(\theta_h + \theta_l)$, this condition amounts to

$$\gamma \leq 2\left(\frac{p}{1-p} + 1\right).$$

Since in this example $\gamma = 2$, this condition is satisfied regardless of q.

We focus on symmetric allocation rules:

$$w_1 = w_2, x_1 = y_2, y_1 = x_2, z_1 = z_2.$$

Note that for the purpose of obtaining a constrained efficient allocation rule, assuming symmetry is in fact without loss.[12] Given this, the problem amounts to

$$\max_{w_1,x_1,y_1,z_1} \frac{1}{4}(2(3z_1 + 31y_1 + 17x_1 + 45w_1))$$

s.t. the feasibility and the **p**-monotonicity. By Lemma 2.2, it is sufficient to consider **p**-monotonicity when the other agent uniformly babbles: when $\psi_2(\theta_2) = \theta_l$ for all θ_2,

$$p\left(\frac{1}{2}x_1 + \frac{1}{2}w_1\right) + (1-p)x_1 \geq p\left(\frac{1}{2}z_1 + \frac{1}{2}y_1\right) + (1-p)z_1 \tag{2.3}$$

and when $\psi_2(\theta_2) = \theta_h$ for all θ_2,

$$p\left(\frac{1}{2}x_1 + \frac{1}{2}w_1\right) + (1-p)w_1 \geq p\left(\frac{1}{2}z_1 + \frac{1}{2}y_1\right) + (1-p)y_1 \tag{2.4}$$

and from the symmetry, the feasibility is now

$$w_1 \leq \frac{1}{2}, z_1 \leq \frac{1}{2}, x_1 + y_1 \leq 1 \tag{2.5}$$

where the last one comes from $x_1 + x_2 \leq 1$ and $y_1 + y_2 \leq 1$.

Proposition 2.4. *The efficient allocation rule which is p-dominant implementable has the following structure:*

(1) When $p \geq \frac{3}{7}$, the optimal solution is $z_1 = 0$, $y_1 = p\frac{3}{4} + (1-p)\frac{1}{2}$ (and $x_1 = 1 - y_1$, $w_1 = 1/2$), and the maximum social surplus is *strictly* increasing in $p \in [3/7, 1]$; in particular, when $p = 1$, it coincides with the most efficient BNE implementable allocation rule.

(2) When $p < 3/7$, $z_1 = 1/2$, $y_1 = 1/2$ ($x_1 = 1/2$, $w_1 = 1/2$), which are the same as the most efficient ex-post implementable allocation rule; the maximum surplus is constant for $p \in [0, 3/7]$.

Proof. See Appendix. □

In this example, the second condition in Theorem 2.1 does not bite. Hence, the constrained allocation is only determined by p-monotonicity. See Figure 2.5.

[12] This is because of the linearity of objective function and the symmetry between agent 1 and agent 2: if there is an asymmetric constrained efficient allocation rule, the the allocation rule in which the role of agent 1 and agent 2 is changed is also constrained efficient. Then, we make the average of the two; and it is still constrained efficient because of the linearity of the objective.

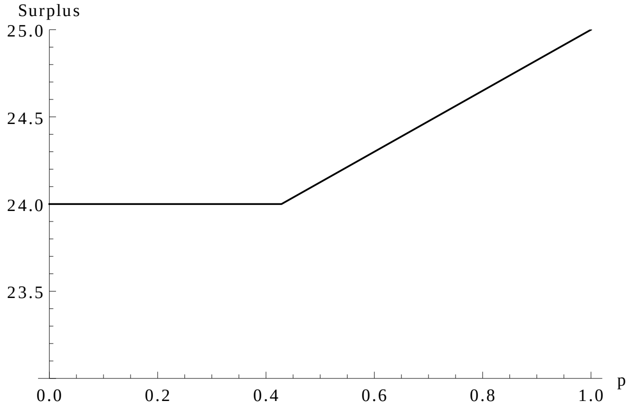


Figure 2.5: The social surplus of the most efficient p-dominant implementable allocation for $p \in [0, 1]$ in Example 2.4

2.6 p-dominant Implementability with Continuous Payoff Type Spaces

In this section, we study **p**-dominant implementability with continuous type spaces. Throughout this section, we assume that there is a probability space and θ is a random variable whose range is $[0, 1]^N$. We further assume θ is absolutely continuous w.r.t. the Lebesgue measure, and let λ be the corresponding density.

First, we generally study **p**-dominant incentive compatibility. Then, we focus on the quasilinear environment, and characterize the set of **p**-dominant allocation rules.

2.6.1 Dichotomy Result

We first look at the private value case.

Proposition 2.5. *With private value and continuous type space, for any* $\mathbf{p}$ *with* $p_i < 1$ *for all* i*, any* $\mathbf{p}$*-dominant implementable social choice correspondence must be dominant strategy implementable.*

Proof. See Appendix. □

In words, there are only two cases, depending on p_i: if $p_i = 1$, then it is the same as BIC incentive compatibility; for any other cases, it requires dominant incentive compatibility. In particular, it is also true when each p_i is arbitrarily close to 1, but not exactly 1. Put differently, there is a discontinuity at $\mathbf{p} = \mathbf{1}$, i.e., as long as an agent believes that the other agents babble with arbitrarily small probability, the result shows that it should be the truth-telling report must be (weakly) dominant.

To understand such a discontinuity, it is enough to observe that for any $p_i < 1$, the incentive compatibility requires that the derivative with respect to i's report involving the babbling part should be the same for every babbling. Then, why the derivative for this part should be 0? This is because, for the babbling part, we also need to consider the babbling that corresponds to the truthful-reporting. Thus, the derivative should be 0.

Note that this result exhibits substantial difference with that for the discrete type case in which we already observe that the set of $\mathbf{p}$-dominant implementable social choice functions may be *strictly* increasing for some $\mathbf{1} > \mathbf{p}' > \mathbf{p}$ (e.g., Example 2.4). Such a difference comes from the fact that we only need inequality for incentive compatibility in the discrete type case, i.e., every babbling does not need to give exactly the same marginal utility; agent i is only required to weakly prefer truthful-reporting for every babbling.

It turns out that a similar account also applies to the interdependent case:

Proposition 2.6. *With interdependent value and continuous type space, for any* $\mathbf{p}$ *with* $p_i < 1$ *for all* i*, any* $\mathbf{p}$*-dominant social choice correspondence is* $\mathbf{0}$*-dominant implementable.*

Proof. See Appendix. □

2.6.2 Characterization of p-dominant Implementable Allocations with quasilinear environment

Based on the previous subsection, it is without loss to only consider dominant implementability and $\mathbf{0}$-dominant implementability for the private value and interdependent value case, respectively. As dominant implementability has been extensively studied in the literature (for example, VCG mechanisms), here we focus on characterizing $\mathbf{0}$-dominant implementability, assuming the interdependent value.

We assume that $q_i : \Theta \to [0,1]$ and $\tau_i : \Theta \to \mathbb{R}$ to be piecewise differentiable w.r.t. θ_i as usually assumed in the literature. Note that most generally they are differentiable almost everywhere, as q_i necessarily satisfies ex-post monotonicity for $\mathbf{0}$-dominant implementability, which will be shown and we can show τ_i is also differentiable at the points q_i is differentiable.[13]

Definition 2.11. An allocation rule $q = (q_i)_{i\in\mathcal{I}}$ satisfies *slope independence* if for each agent $i \in \mathcal{I}$, if q_i is differentiable w.r.t. θ_i at some $\hat{\theta}_i$, then

$$\left.\frac{\partial q_i(\theta_i, \theta_{-i})}{\partial \theta_i}\right|_{\theta_i=\hat{\theta}_i}$$

does not depend on θ_{-i}.

Theorem 2.2. *Suppose $\gamma > 0$. An allocation rule $q = (q_i)_{i\in\mathcal{I}}$ is $\mathbf{0}$-dominant implementable if and only if q satisfies ex-post monotonicity and slope independence.*

We prove this by a sequence of the propositions and lemmas. We have already shown that the monotonicity is necessary (Lemma 2.3). Proposition 2.7 is about the sufficiency; and Proposition 2.8 and Lemma 2.5 together gives the necessity.

Before we provide a formal proof for this result, it may be useful to provide an informal account behind it. In fact, the intuitions are similar to the discrete type case (i.e., Theorem 2.1). Recall that in the discrete type case, the marginal benefit of reporting a higher type depends on the babbling; the difference of the marginal benefit between high type and low type also depends on it. The impact of this can be divided into the private value part and the interdependent part, and the latter depends on the true type of the other agents; that is, the former part only depends on the report of other agents, while the latter depends on which type reports which type, i.e., babbling. The availability of a transfer rule that yields the marginal cost that can be between the marginal benefit for all the babbling, the impact of babbling on the interdependent value should be small enough so that it can be covered by the private part. As the difference of the marginal benefit between two close

[13] Recall that a monotone function is differentiable almost everywhere.

types, the difference of the marginal benefit of this part becomes close too; eventually this part becomes 0. On the other hand, the babbling effect on the interdependent part remains strictly positive.

We can also think directly the continuous type. The marginal benefit and cost together are represented by the derivative of the expected payoff w.r.t. agent i's report. For any babbling, the derivative should be the same as 0; in particular, for any two babblings that induce the same distribution over the other agents', the derivative should be the same. However, two different babblings typically induce different expected marginal benefits, just because the interdependent part in payoff also depends on the other agents true type profile. In turn, this implies that in order to make a mechanism **0**-dominant implementable, it is necessary that different babbling does not yields different derivative; and a sufficient condition for this is the derivative of allocation does not depend on others' report.

Proposition 2.7 (sufficiency). *Suppose q satisfies slope independence. Then, if q satisfies ex-post monotonicity, then it is* **0***-dominant implementable.*

Proof. See Appendix. □

Let us turn to the necessity.

Proposition 2.8. *Let $q = (q_i)_i$ be an allocation rule and* **0***-dominant implementable. Then, for all $i \in \mathcal{I}$, θ_i, $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$,*

$$\gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \mu_j \right) \left. \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \right|_{\theta_i' = \theta_i} \right] = \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \left. \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \right|_{\theta_i' = \theta_i} \right]. \quad (2.6)$$

Proof. See Appendix. □

Lemma 2.5. *Assume $\gamma > 0$. Suppose that q does not satisfy slope independence. That is, there exist some $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_{-i}, \theta_{-i''} \in \Theta_{-i}$ such that*

$$\frac{\partial q_i(\theta_i, \theta''_{-i})}{\partial \theta_i} > \frac{\partial q_i(\theta_i, \theta'_{-i})}{\partial \theta_i}.$$

Then, condition (2.6) is violated.

Proof. See Appendix. □

The following is a corollary of the theorem,

Corollary 2.6. *In this environment, any* $\mathbf{p}$*-dominant implementable allocation rule where* $p_i < 1$ *for all* $i \in \mathcal{I}$ *is ex-post implementable.*

2.6.3 Constrained Efficient 0-dominant Implementable Allocation Rule

We have the following observation immediately from Theorem 2.2 and Proposition 2.8.

Proposition 2.9. *Suppose* $\gamma > 0$ *and* $\gamma \neq 1$*. Then, the efficient allocation rule is not* $\mathbf{p}$*-dominant implementable for any* $\mathbf{p}$ *such that* $p_i < 1$ *for all* $i \in \mathcal{I}$*.*

Note that, by Proposition 2.6, this means that for any $\mathbf{p}$ where $p_i < 1$ for any $i \in \mathcal{I}$, the efficient allocation is not $\mathbf{p}$-dominant implementable. This might come as a surprise, because we know that when $\mathbf{p} = \mathbf{1}$, the efficient allocation is implementable as long as $\gamma < 1$ (Maskin, 1992) through what he called the generalized VCG mechanism. The result says that this positive result is not robust in the sense that it is not $\mathbf{p}$-dominant implementable even for $\mathbf{p}$ arbitrarily close to $\mathbf{1}$.

Given this result, a natural important question is what is the most efficient allocation rule that is $\mathbf{0}$-dominant implementable. In the rest of this subsection, we study this question. In doing so, we also show the distinctive feature of slope independent allocation rules.

To study this question, a natural starting point is to consider (potentially random) *constant* allocation rules; namely, no agents affect the allocation. Obviously, they are implementable in dominant strategy equilibrium, so $\mathbf{0}$-dominant implementable.

Example 2.5 (Constant random allocation is not constrained efficient)**.** It will be shown shortly that a constant random allocation rule (i.e., giving the item randomly to agents) is not constrained efficient; Suppose that there are two agents $i = 1, 2$ and their payoff types are i.i.d drawn from the uniform distribution on $[0, 1]$. For simplicity, let us assume that the item will be given to an agent with equal probability. Obviously, it is $\mathbf{0}$-dominant implementable. In this case, the corresponding ex-ante expected social surplus is

$$\mathbb{E}\left[\frac{1}{2}(\theta_1+\gamma\theta_2)+\frac{1}{2}(\theta_2+\gamma\theta_1)\right]=\frac{1}{2}+\gamma\frac{1}{2}. \tag{2.7}$$

To show that the the constrained inefficiency of this allocation rule, consider instead the following allocation rule,

$$\begin{aligned} q_1(\theta_1,\theta_2) &= \theta_1 \\ q_2(\theta_1,\theta_2) &= 1-q_1(\theta_1,\theta_2)=1-\theta_1 \end{aligned}$$

As q satisfies slope independence and it satisfies the monotonicity, it is **0**-dominant implementable by the above proposition. In this case, the social surplus is

$$\begin{aligned} \mathbb{E}[\theta_1(\theta_1+\gamma\theta_2)]+\mathbb{E}[(1-\theta_1)(\theta_2+\gamma\theta_1)] &= \mathbb{E}[\theta_1^2+\gamma\theta_1\theta_2]+\mathbb{E}[(1-\theta_1)\theta_2+\gamma(1-\theta_1)\theta_1] \\ &= \frac{1}{3}+\gamma\frac{1}{4}+\frac{1}{2}\frac{1}{2}+\gamma\left(\frac{1}{2}-\frac{1}{3}\right) \\ &= \frac{7}{12}+\gamma\frac{5}{12}. \end{aligned}$$

Compare this to the value in (2.7).

$$\frac{1}{2}+\gamma\frac{1}{2}<\frac{7}{12}+\gamma\frac{5}{12} \iff \gamma<1.$$

That is, if $\gamma<1$, then the latter allocation rule is more efficient.

We can “symmetrize” this mechanism as follows: with probability $1/2$, approach to agent 1 and exercise the above mechanism, and with the rest probability approaches to agent 2 and do similarly. The resulting mechanism is

$$\begin{aligned} q_1(\theta_1,\theta_2) &= \frac{1}{2}\theta_1+\frac{1}{2}(1-\theta_2)=\frac{1}{2}+\frac{1}{2}(\theta_1-\theta_2) \\ q_2(\theta_1,\theta_2) &= \frac{1}{2}+\frac{1}{2}(\theta_2-\theta_1). \end{aligned}$$

2.6.3.1 Asking only one agent

A meaningful observation is “asking only one agent” mechanism (more precisely, allocation rule) like the above is **0**-dominant implementable.

Corollary 2.7. *Any allocation rule that only depends on one and only one agent's report and satisfies ex-post monotonicity is* **0**-*implementable.*

Is the above mechanism most efficient in this class? It turns out that there is more efficient allocation rule that is **0**-dominant implementable: fix an agent, say agent 1, and approach only to agent 1 and ask his/her type and if it is bigger than the other agent's mean, then give the item to agent 1; otherwise give it to agent 2. Let us see whether it gives a higher social surplus:

$$\begin{aligned}\mathbb{E}[(\theta_1+\gamma\theta_2)\mathbf{1}_{\{\theta_1>1/2\}}+(\theta_2+\gamma\theta_1)\mathbf{1}_{\{\theta_1\le 1/2\}}] &= \int_0^{1/2}(1/2+\gamma\theta_1)d\theta_1+\int_{1/2}^{1}(\theta_1+\gamma 1/2)d\theta_1\\ &= \frac{5}{8}+\gamma\frac{3}{8}.\end{aligned}$$

Note that

$$\frac{5}{8}+\gamma\frac{3}{8}>\frac{7}{12}+\gamma\frac{5}{12}$$

as long as $\gamma>1$. In fact we will show that this allocation rule is most efficient among the "asking only one agent" **0**-dominant implementable allocation rules.

Proposition 2.10. *Suppose $N=2$ and $\gamma\in(0,1)$. Fix $i\in\mathcal{I}$. Suppose an allocation rule $q=(q_i)_i$ only depends on θ_i. Then, the following allocation rule is the most efficient* **0**-*dominant allocation rule under this assumption:*

$$\begin{aligned}q_i(\theta)&=\mathbf{1}_{\{\theta_i>\mu_j\}}\\ q_j(\theta)&=1-q_i(\theta)=\mathbf{1}_{\{\theta_i\le\mu_j\}}\end{aligned}$$

where $j\neq i$. In this case, the ex-ante expected social surplus is

$$V^i:=\mu_j+\gamma\mu_i+\int_{\theta_i>\mu_j}((\theta_i+\gamma\mu_j)-(\mu_j+\gamma\theta_i))\lambda_i(\theta_i)d\theta_i \tag{2.8}$$

Proof. Since allocation rule does not depend on θ_j where $j\neq i$, without loss of generality, we can assume that

$$q_i(\theta)=y(\theta_i)$$

and due to efficiency

$$q_j(\theta)=1-y(\theta_i)$$

where $y : \Theta_i \to [0,1]$ and increasing. The corresponding expected social surplus is

$$\begin{aligned}\mathbb{E}[(\theta_i + \gamma\theta_j)y(\theta_i) + (\theta_j + \gamma\theta_i)(1 - y(\theta_i))] &= \mathbb{E}[(\theta_i + \gamma\mu_j)y(\theta_i) + (\mu_j + \gamma\theta_i)(1 - y(\theta_i))] \\ &= \mu_j + \gamma\mu_i + \int_0^1 ((\theta_i + \gamma\mu_j) - (\mu_j + \gamma\theta_i))y(\theta_i)\lambda_i(\theta_i)d\theta_i.\end{aligned}$$

Note that under our assumption $\gamma < 1$

$$\theta_i + \gamma\mu_j > \mu_j + \gamma\theta_i \iff \theta_i > \mu_j$$

As $y(\theta_i) \in [0,1]$, the pointwise maximization gives that the optimal $y(\theta_i)$ is 1 if $\theta_i > \mu_j$; otherwise 0. Note that it is increasing in θ_i as desired. □

From this proposition, we now know that what is the most efficient allocation rule that is $\mathbf{0}$-dominant implementable among "asking only one agent":

Corollary 2.8. *The most efficient* $\mathbf{0}$*-dominant allocation rule among "asking only one agent" can be obtained by maximizing the social surplus that corresponds to the mechanism in the proposition over* i*.*

Intuitively, the mechanism that only asks agent i can be interpreted as if the designer has the opportunity cost which is the same as the mean of $\theta_j + \gamma\mu_i$, $j \neq i$.

2.6.3.2 General mechanism

Proposition 2.11 (Efficient $\mathbf{0}$-dominant allocation rule when $N = 2$)**.** *Assume* $N = 2$*. Then, the constrained efficient* $\mathbf{0}$*-dominant implementable allocation rule is*

$$\begin{aligned}q_i(\theta) &= \mathbf{1}_{\{\theta_i > \mu_j\}} \\ q_j(\theta) &= 1 - q_i(\theta) = \mathbf{1}_{\{\theta_i \le \mu_j\}}\end{aligned}$$

where $i \in \mathcal{I}$ *is such that*

$$i \in \underset{i \in \{1,2\}}{\arg\max}\, V^i$$

where V^i *is the expression in (2.8).*

Proof. See Appendix. □

2.7 Discussion: Formal Robustness Foundation of p-dominant Implementation

2.7.1 Robust Foundations for p-dominant Implementation

We find that there are some social choice correspondences (especially functions) that *are* ex-post implementable, but not **p**-dominant implementable in some type space with some $\mathbf{p} \in [0,1]^N$. This implies (by Corollary 1 of Bergemann and Morris (2005)) that there is some social choice function that is implementable in BNE in any (non redundant) type spaces; nevertheless, it is not **p**-dominant implementable. On the one hand, this suggests that **p**-dominant implementability is unnecessarily strong to capture robustness to uncertainty in information structure. On the other hand, it suggests that **p**-dominant implementability may capture "additional" robustness beyond payoff environment. In this subsection, we provide two additional robustness concern that **p**-dominant implementability may capture.

2.7.2 Irrational Agents

Eliaz (2002) studies situations where some agents might be *faulty* in the sense that they do not choose their action according to incentives. His approach is to study mechanisms that are immune to the possibility of at most $k \leq N$ faulty agents. Here we take a different approach. We call an agent *irrational* if he is faulty in the sense of Eliaz. We assume that if an agent is irrational, he/she chooses some arbitrary action potentially mixed.[14]

Consider the following type space. $\mathcal{T}_r \equiv ((S_i)_{i\in\mathcal{I}}, (\beta_i)_{i\in\mathcal{I}})$ where

$$S_i = \{r\} \cup M_i, \forall i \in \mathcal{I}$$

where M_i is the message space for agent i. Here $s_i = r$ represents *rational* type and $s_i = a_i$ represents *irrational* type who is supposed to play $a_i \in M_i$ regardless of θ.[15]

[14] In some context (e.g., some reputation models), such irrationality is captured by a particular preference; however, with interdependent value, information about payoff-relevant parameter, i.e., payoff type, seems conceptually different from our notion of irrationality.

[15] One may extend or reduce it, i.e., including some mixed action profile, still holding S_i countable.

Note that by this modeling, we allow that rationality of agents are correlated. Define

$$S^r \equiv \{(r, \dots, r)\}.$$

That is, the event that every agent is rational (and each agent knows he is rational). The standard mechanism design problem implicitly or explicitly assume that every agent is rational and it is common knowledge; this can be modeled as $S = \{(r, \dots, r)\}$. Here we relax this assumption.

Given a payoff type space $((\Theta_i)_{i\in\mathcal{I}}, (\beta_i)_{i\in\mathcal{I}})$, define

$$T_i := \Theta_i \times S_i$$
$$\pi_i(\theta_{-i}, s_{-i}|\theta_i, s_i) := \lambda_i(\theta_{-i}|\theta_i)\beta_i(s_{-i}|s_i).$$

Let us modify the definition of Bayes Nash equilibrium to accommodate irrational agents so that we do not concern about irrational agents' incentive compatibility.

Proposition 2.12. *Let $\mathcal{M} = (M, g)$ be a mechanism and $\sigma \equiv (\sigma_i)_{i\in\mathcal{I}}$ be a $\mathbf{p}$-dominant equilibrium in $(\mathcal{M}, ((\Theta_i)_{i\in\mathcal{I}}, \lambda))$. Consider $\mathcal{T}_r = (S, (\beta_i)_i)$ defined above. Suppose $S^r = \{(r, \dots, r)\}$ is $\mathbf{p}$-evident event. Then there is a BNE, $\sigma' \equiv (\sigma'_i)_{i\in\mathcal{I}}$ of the auxiliary game with $((T_i)_i, (\pi_i)_i)$ such that*

$$\sigma'_i(r, \theta_i) = \sigma_i(\theta_i), \forall i \in \mathcal{I}, \theta_i \in \Theta_i.$$

Proof. In order to construct $\sigma' \equiv (\sigma'_i)_{i\in\mathcal{I}}$, consider a modified game in which every agent i with type (θ_i, r) is fixed to play $\sigma_i(\theta_i)$. In this Bayesian game, we know that there exists an Bayes Nash equilibrium. Let us call this $\sigma'' \equiv (\sigma''_i)_{i\in\mathcal{I}}$. For each i, construct σ'_i as follows:

$$\sigma'_i(\theta_i, r) := \sigma_i(\theta_i), \forall \theta_i \in \Theta_i$$

and for any $s_i \neq r$,

$$\sigma'_i(\theta_i, s_i) := \sigma''_i(\theta_i, s_i), \forall \theta_i \in \Theta_i.$$

By construction, for any type (θ_i, s_i) with $s_i \neq r$ with $s_i = r$, the incentive compatibility is satisfied. To see this for (θ_i, r), note that

$$\begin{aligned}
&\sum_{s_{-i}} \beta_i(s_{-i}|s_i) \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m'_i, \sigma'_{-i}(s_{-i}, \theta_{-i})), \theta_i, \theta_{-i}) \\
&\quad = \sum_{s_{-i}=\mathbf{r}_{-i}} \beta_i(s_{-i}|s_i) \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m'_i, \sigma_{-i}(s_{-i}, \theta_{-i})), \theta_i, \theta_{-i}) \\
&\qquad + \sum_{s_{-i}\neq\mathbf{r}_{-i}} \beta_i(s_{-i}|s_i) \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m'_i, \sigma''_{-i}(s_{-i}, \theta_{-i})), \theta_i, \theta_{-i}) \\
&\quad = \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) \left(\sum_{s_{-i}=\mathbf{r}_{-i}} \beta_i(s_{-i}|s_i) \sigma_{-i}(\theta_{-i}) + \sum_{s_{-i}\neq\mathbf{r}_{-i}} \beta_i(s_{-i}|s_i) \sigma_{-i}(\theta_{-i}, s_{-i}) \right) \\
&\qquad \times u_i(g(m'_i, m_{-i}), \theta_i, \theta_{-i}) \\
&\quad = \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m'_i, q_i^{s_i} \sigma_{-i}(\theta_{-i}) + (1 - q_i^{s_i}) \phi_{-i}^{s_i}(\theta_{-i})), \theta_i, \theta_{-i})
\end{aligned}$$

where

$$\begin{aligned}
q_i^{s_i} &\equiv \sum_{s_{-i}\in S^*_{-i}} \beta_i(\omega, s_{-i}|s_i) \\
\phi_{-i}^{s_i}(\theta_{-i}) &\equiv \frac{1}{1 - q_i^{s_i}} \sum_{s_{-i}\in S_{-i}\setminus S^*_{-i}} \beta_i(s_{-i}|s_i) \sigma_{-i}(\theta_{-i}, s_{-i}).
\end{aligned}$$

Note that $\phi_{-i}^{s_i}(\theta_{-i})$ is a distribution over M_{-i}.[16] Note that $q_i^{s_i} \geq p_i$, because S^* is $\mathbf{p}$-evident. By definition of $\mathbf{p}$-dominant equilibrium,

$$m_i \in \arg\max_{m'_i \in M_i} \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m'_i, q_i^{s_i} \sigma_{-i}(\theta_{-i}) + (1 - q_i^{s_i}) \phi_{-i}^{s_i}(\theta_{-i})), \theta_i, \theta_{-i}).$$

□

To provide a version of the converse, we slightly extend $\mathcal{T}_r = (S, \beta_i)$ in the following way; we extend β_i so that $\beta_i : \Theta_i \times S_{-i} \to \Delta(\Theta_{-i} \times S_{-i})$ such that for all i, θ_i and s_i

$$\sum_{s_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, s_i) = \lambda_{-i}(\theta_{-i}|\theta_i), \forall \theta_{-i}. \tag{2.9}$$

[16] To see this,

$$\frac{1}{1 - q_i s_i} \sum (\omega, s_{-i}) \in \Omega \times (S_{-i} \setminus S^*_{-i}) \beta_i(\omega, s_{-i}|s_i) \sum_{m_{-i}} \sigma_{-i}(\theta_{-i}, s_{-i})(m_{-i}) = \frac{1}{1 - q_i s_i}(1 - q_i s_i) = 1.$$

Namely, s_i and θ_i may be correlated, but observing s_i does not give further information about θ_{-i} given θ_i.

Proposition 2.13. *Let $\mathbf{p} \in [0,1]^N$ and $\mathcal{M} \equiv (M, g)$ be a mechanism, and suppose a BNE $\sigma \equiv (\sigma_i)_{i\in\mathcal{I}}$ in $(\mathcal{M}, ((\Theta_i)_{i\in\mathcal{I}}, \lambda))$ is not a $\mathbf{p}$-dominant equilibrium. Then, there exists $\mathcal{T}_r = (S, (\beta_i)_{i\in\mathcal{I}})$ such that the event $\Theta \times S^r$ is common $\mathbf{p}$-belief and there is no equilibrium such that when $s \in S^r$, σ is played for every θ.*

In words, if a BNE $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ is not a $\mathbf{p}$-dominant equilibrium, then we can find some information structure in which the rationality of each agent is common $\mathbf{p}$-believed; nevertheless, σ does not satisfy Bayesian incentive compatibility for some realization of θ_i for some agent i given that the other agent plays according to σ_j as far as they are rational.

Proof. Since σ is not a $\mathbf{p}$-dominant equilibrium, there exists $i \in \mathcal{I}$, $\theta_i \in \Theta_i$, $m_i \in M_i$ with $\sigma_i(\theta_i)[m_i] > 0$, $m_i' \in M_i$ such that

$$\sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m_i', \phi_{-i}(\theta_{-i}), \theta_i, \theta_{-i}) > \sum_{\theta_{-i}} u_i(g(m_i, \phi_{-i}(\theta_{-i})), \theta_i, \theta_{-i}) \tag{2.10}$$

for some $\phi_{-i} : \Theta_{-i} \to \Delta(M_{-i})$ such that $\phi_{-i}(\theta_{-i}) = q_i^{\theta_{-i}} \sigma_i(\theta_{-i}) + (1 - q_i^{\theta_{-i}})\psi_{-i}(\theta_{-i})$ where $q_i^{\theta_{-i}} \geq p_i$ for any θ_{-i}.

We construct $\mathcal{T}_r$ as follows: for each θ_{-i},

$$\begin{aligned} \beta_i(\theta_{-i}, \mathbf{r}_{-i}|\theta_i, r) &= q_i^{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) \\ \beta_i(\theta_{-i}, m_{-i}|\theta_i, r) &= (1 - q_i^{\theta_{-i}})\psi_{-i}(\theta_{-i})[m_{-i}]\lambda_i(\theta_{-i}|\theta_i). \end{aligned}$$

Note that

$$\begin{aligned} \sum_{s_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, r) &= q_i^{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) + (1 - q_i^{\theta_{-i}}) \left(\sum_{m_{-i}} \psi_{-i}(\theta_{-i})[m_{-i}] \right) \lambda_i(\theta_{-i}|\theta_i) \\ &= \lambda_i(\theta_{-i}|\theta_i). \end{aligned}$$

For all other $(\theta_i', s_i) \neq (\theta_i, r)$ and for $j \neq i$ define $\beta_j(\theta_{-j}, s_{-j}|\theta_j, s_j) = \lambda_j(\theta_{-i}|\theta_j)$ when $s_{-j} = (r, \ldots, r)$; otherwise 0.

Note that $S^r = (r, \ldots, r)$ is **p**-evident by observing

$$\beta_i(\Theta_{-i} \times \{\mathbf{r}_{-i}\}|\theta_i, r) \geq \min_{\theta_{-i}} q_i^{\theta_{-i}} \geq p_i$$

and for any $j \in \mathcal{I}$ and $(\theta_j, r) \neq (\theta_i, r)$,

$$\beta_j(\Theta_{-i} \times \{\mathbf{r}_{-j}\}|\theta_j, r) = 1.$$

Consider any strategy $\sigma' \equiv (\sigma'_i)_{i \in \mathcal{I}}$ in $\mathcal{T}_r$ which satisfies $\sigma'_j(\theta_j, s_j) = \sigma_j(\theta_j)$ for $s_j \neq s^*_j$; otherwise arbitrary. We need to check sending m_i is not incentive compatible for i with θ_i. Let $s_i = r$.

$$\begin{aligned}
&\sum_{\theta_{-i}, s_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, r) u_i(g(m_i, \sigma'_{-i}(\theta_{-i}, s_{-i})), \theta_i, \theta_{-i}) \\
&= \sum_{\theta_{-i}, s_{-i} : s_{-i} = \mathbf{r}_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, r) u_i(g(m_i, \sigma'_{-i}(\theta_{-i}, s_{-i})), \theta_i, \theta_{-i}) \\
&\quad + \sum_{\theta_{-i}, s_{-i} : s_{-i} \neq \mathbf{r}_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, r) u_i(g(m_i, \sigma'_{-i}(\theta_{-i}, s_{-i})), \theta_i, \theta_{-i}) \\
&= \sum_{\theta_{-i}, s_{-i} : s_{-i} = \mathbf{r}_{-i}} \beta_i(\theta_{-i}, s_{-i}|\theta_i, r) u_i(g(m_i, \sigma_{-i}(\theta_{-i})), \theta_i, \theta_{-i}) \\
&\quad + \sum_{\theta_{-i}, m_{-i}} \beta_i(\theta_{-i}, m_{-i}|\theta_i, r) u_i(g(m_i, m_{-i}), \theta_i, \theta_{-i}) \\
&= \sum_{\theta_{-i}, s_{-i} : s_{-i} = \mathbf{r}_{-i}} q_i^{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) u_i(g(m_i, \sigma_{-i}(\theta_{-i})), \theta_i, \theta_{-i}) \\
&\quad + \sum_{\theta_{-i}, m_{-i}} (1 - q_i^{\theta_{-i}}) \psi_{-i}(\theta_{-i})[m_{-i}] \lambda_i(\theta_{-i}|\theta_i) u_i(g(m_i, m_{-i}), \theta_i, \theta_{-i}).
\end{aligned}$$

Thus, by (2.10), sending m'_i is a profitable deviation. □

2.8 Conclusion

In this paper, we completely characterize **p**-dominant implementability with payoff type spaces and quasilinear environment. In particular, we find an extra condition other than the monotonicity for an allocation to be **p**-dominant implementable. Interestingly, **p**-dominant implementability and ex-post implementability do not generally imply each other. This suggests that **p**-dominant implementability may capture robustness to uncertainty of payoff-environment. We also provide a formal robustness foundation for **p**-dominant implementability.

There are some future directions which we believe worth working on. **p**-dominant implementable we study in this paper is a partial implementation concept, but still requiring the interdependence condition. To best our knowledge, this condition does not appear in other research on robust partial implementation (e.g., Bergemann and Morris (2005)). As mentioned, a version of the condition can be found in full implementation literature (e.g., Bergemann and Morris (2009a); Ollár and Penta (2017)). It would be interesting to clarify this point.

2.9 Appendix

2.9.1 Omitted Proofs

2.9.1.1 Proof of Lemma 2.1

The necessity part is trivial as they are a subset if ICs needs to be checked.

For the sufficiency part, consider incentive of agent i with $t_i \in T_i$ when report t'_i when the other agents babble $\phi_{-i}(t_{-i}) \in \Delta(T_{-i})$

$$\sum_{t_{-i}} \lambda_i(t_{-i}|t_i) \sum_{t'_{-i}} \phi_{-i}(t'_{-i}|t_{-i}) u_i(f(\tilde{\theta}_i(t'_i), \tilde{\theta}_{-i}(t'_{-i})), \tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})).$$

For notational convenience, let $\tilde{f}(t_i, t_{-i}) \equiv f(\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$. By definition of **p**-dominant equilibrium, $\phi_{-i}(t_{-i})[t_{-i}] \geq p_i$,

$$p_i \left(\sum_{t_{-i}} \lambda_i(t_{-i}|t_i) u_i(\tilde{f}(t'_i, t_{-i}), t_i, t_{-i}) \right) \\ + (1 - p_i) \sum_{t_{-i}} \lambda_i(t_{-i}|t_i) \sum_{t'_{-i}} \tilde{\phi}_{-i}(t'_{-i}|t_{-i}) u_i(\tilde{f}(t'_i, t'_{-i})), t_i, t_{-i}) \quad (2.11)$$

where

$$\tilde{\phi}_{-i}(t_{-i})[t'_{-i}] := \frac{1}{1-p_i}\phi_{-i}(t_{-i})[t'_{-i}], \forall t'_{-i} \neq t_{-i}$$
$$\tilde{\phi}_{-i}(t_{-i})[t_{-i}] := \frac{1}{1-p_i}(\phi_{-i}(t_{-i})[t_{-i}] - p_i) \geq 0.$$

Then note that

$$\begin{aligned}\sum_{t'_{-i}\in T_{-i}} \tilde{\phi}_{-i}(t_{-i})[t'_{-i}] &= \sum_{t'_{-i}\neq t_{-i}} \frac{1}{1-p_i}\phi_{-i}(t_{-i})[t'_{-i}] + \frac{1}{1-p_i}(\phi_{-i}(t_{-i})[t_{-i}] - p_i) \\ &= \sum_{t'_{-i}\in T_{-i}} \frac{1}{1-p_i}\phi_{-i}(t_{-i})[t'_{-i}] + \frac{-p_i}{1-p_i} = \frac{1}{1-p_i} + \frac{-p_i}{1-p_i} = 1.\end{aligned}$$

That is, given t_{-i}, the last term of (2.11) is a convex combination of $u_i(\tilde{f}(t'_i, t'_{-i}), t_i, t_{-i})$ with the weight $\tilde{\phi}_{-i}(t_{-i})[t'_{-i}]$. Thus,

$$\begin{aligned} t_i \in \arg\max_{t'_i \in t_i} p_i &\left(\sum_{t_{-i}} \lambda_i(t_{-i}|t_i) u_i(\tilde{f}(t'_i, t_{-i}), t_i, t_{-i}) \right) \\ &+ (1-p_i) \sum_{t_{-i}} \lambda_i(t_{-i}|t_i) u_i(\tilde{f}(t'_i, \psi_{-i}(t_{-i})), t_i, t_{-i}), \forall \psi_{-i} : T_{-i} \to T_{-i} \end{aligned}$$

implies

$$\begin{aligned} t_i \in \arg\max_{t'_i \in t_i} p_i &\left(\sum_{t_{-i}} \lambda_i(t_{-i}|t_i) \sum_{t'_{-i}} u_i(\tilde{f}(t'_i, t'_{-i}), t_i, t_{-i}) \right) \\ &+ (1-p_i) \sum_{t_{-i}} \lambda_i(t_{-i}|t_i) \sum_{t'_{-i}} \tilde{\phi}_{-i}(t'_{-i}|t_{-i}) u_i(\tilde{f}(t'_i, t'_{-i}), t_i, t_{-i}). \end{aligned}$$

2.9.1.2 Proof of Lemma 2.2

"Only if" part is trivial, as uniform babbling is subset of babbling.

For "if" part, given any function $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$, note that

$$Q_i^{p_i}(\theta_i, \psi_{-i}) = \sum_{\theta_{-i}} \lambda_i(\theta_{-i}|\theta_i) Q_i^{p_i}(\theta_i, \psi_{-i}(\theta_{-i}))$$

Thus, if $Q_i^{p_i}(\theta_i, \theta_{-i})$ is monotone in θ_i for any θ_{-i}, then $Q_i^{p_i}(\theta_i, \psi_{-i})$ is also monotone in θ_i.

2.9.1.3 Proof of Proposition 2.1

Consider agent the incentive compatibility of agent i with type t_i. By definition of **p**-dominant equilibrium,

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(\sigma_i(t_i), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$
$$\geq \sum_{t_{-i}} \beta_i(t_{-i}|t_i)u_i(g(m_i', \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall m_i' \in M_i$$

for any $\phi_{-i}(t_{-i}) \in \Delta(M_{-i})$ such that each player $j \neq i$ with t_j plays $\sigma_j(t_j)$ with probability at least p_i and arbitrarily with the rest probability. In particular,

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(\sigma_i(t_i), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$
$$\geq \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(\sigma_i(t_i'), \phi_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i' \in T_i$$

This implies

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(\sigma_i(t_i), \phi'_{-i}(t_{-i})), \tilde{\theta}_i t_i, \tilde{\theta}_{-i}(t_{-i}))$$
$$\geq \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(g(\sigma_i(t_i'), \phi'_{-i}(t_{-i})), \tilde{\theta}(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i \in T_i'$$

where $\phi'_{-i}(t_{-i}) \in \Delta(M_{-i})$ where each $j \neq i$ plays $\sigma_j(t_j)$ with at least probability p_i and plays arbitrarily m_j from

$$M_j' \equiv \{m_j \in M_j : \exists t_j \in T_j, \sigma_j(t_j)[m_j] > 0\}.$$

That is, the set of messages which are sent by some t_j' in σ_j with positive probability.

Define

$$f(t_i, t_{-i}) := g(\sigma_i(t_i), \sigma_{-i}(t_{-i})), \forall t_i \in T_i, t_{-i} \in T_{-i}.$$

Then the above inequality implies

$$\sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(f(t_i, \tilde{\phi}_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i}))$$
$$\geq \sum_{t_{-i}\in T_{-i}} \beta_i(t_{-i}|t_i)u_i(f(t_i', \tilde{\phi}_{-i}(t_{-i})), (\tilde{\theta}_i(t_i), \tilde{\theta}_{-i}(t_{-i})), \forall t_i' \in T_i \quad (2.12)$$

for any $\tilde{\phi}_{-i}(t_{-i})$ such that any $j \neq i$ reports truthfully at least probability p_i. Lastly, observe that

$$f(t) = g(\sigma(t)) \in F(\tilde{\theta}(t)).$$

2.9.1.4 Proof of Lemma 2.3

Consider agent i's incentive compatibility when the opponents employ $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$ as their babbling. Let $\theta_i' > \theta_i$ and suppose

$$Q_i^{p_i}(\theta_i, \psi_{-i}) > Q_i^{p_i}(\theta_i', \psi_{-i}). \tag{2.13}$$

Consider the incentive compatibility of agent i with type θ_i.

$$\begin{aligned} & p_i \mathbb{E}_{\theta_{-i}} \left[\left(\theta_i + \gamma \sum_{j \neq i} \theta_j \right) q_i(\theta_i, \theta_{-i}) \right] + (1-p_i) \mathbb{E}_{\theta_{-i}} \left[\left(\theta_i + \gamma \sum_{j \neq i} \theta_j \right) q_i(\theta_i, \psi_{-i}(\theta_{-i})) \right] - T_i^{p_i}(\theta_i, \psi_{-i}) \\ \geq\ & p_i \mathbb{E}_{\theta_{-i}} \left[\left(\theta_i + \gamma \sum_{j \neq i} \theta_j \right) q_i(\theta_i', \theta_{-i}) \right] + (1-p_i) \mathbb{E}_{\theta_{-i}} \left[\left(\theta_i + \gamma \sum_{j \neq i} \theta_j \right) q_i(\theta_i', \psi_{-i}(\theta_{-i})) \right] - T_i^{p_i}(\theta_i', \psi_{-i}) \end{aligned}$$

where

$$T_i^{p_i}(\theta_i', \psi_{-i}) \equiv p_i \mathbb{E}_{\theta_{-i}}[\tau_i(\theta_i', \theta_{-i})] + (1-p_i)\mathbb{E}_{\theta_{-i}}[\tau_i(\theta_i', \psi_{-i}(\theta_{-i}))]$$

This can be written as follows:

$$\begin{aligned} & \theta_i(Q_i^{p_i}(\theta_i, \psi_{-i}) - Q_i^{p_i}(\theta_i', \psi_{-i})) \\ & + \mathbb{E}_{-i} \Bigg[\gamma \left(\sum_{j \neq i} \theta_j \right) (p_i(q_i(\theta_i, \theta_{-i}) - q_i(\theta_i', \theta_{-i})) + (1-p_i)(q_i(\theta_i, \psi_{-i}(\theta_{-i}) - q_i(\theta_i', \psi_{-i}(\theta_{-i}))) \Bigg] \\ & \qquad - (T_i^{p_i}(\theta_i, \psi_{-i}) - T_i^{p_i}(\theta_i', \psi_{-i})) \geq 0 \end{aligned}$$

Note that the second term does not involve θ_i. By (2.13), as $\theta_i' > \theta_i$,

$$\begin{gathered} \theta_i'(Q_i^{p_i}(\theta_i, \psi_{-i}) - Q_i^{p_i}(\theta_i', \psi_{-i})) \\ + \mathbb{E}_{-i} \Bigg[\gamma \left(\sum_{j \neq i} \theta_j \right) (p_i(q_i(\theta_i, \theta_{-i}) - q_i(\theta_i', \theta_{-i})) + (1-p_i)(q_i(\theta_i, \psi_{-i}(\theta_{-i}) - q_i(\theta_i', \psi_{-i}(\theta_{-i}))) \Bigg] \\ -(T_i^{p_i}(\theta_i, \psi_{-i}) - T_i^{p_i}(\theta_i', \psi_{-i})) > 0 \end{gathered}$$

That is, θ_i' has incentive to deviate to reporting θ_i. Contradiction.

2.9.1.5 Proof of Theorem 2.1

We introduce following notations:

- Given θ_{-i}, $\Delta q_i(\theta_i^k, \theta_{-i}) \equiv q_i(\theta_i^k, \theta_{-i}) - q_i(\theta_i^{k-1}, \theta_{-i})$, for all $k \geq 1$
- Given θ_{-i}, $\Delta \tau_i(\theta_i^k, \theta_{-i}) \equiv \tau_i(\theta_i^k, \theta_{-i}) - \tau_i(\theta_i^{k-1}, \theta_{-i})$ for all $k \geq 1$
- Given a babbling $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$,

$$\Delta Q_i(\theta_i^k, \psi_{-i}) \equiv \mathbb{E}_{\theta_{-i}}[\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i})]$$

$$\Delta T_i(\theta_i^k, \psi_{-i}) \equiv \mathbb{E}_{\theta_{-i}}[\Delta \tau_i(\theta_i^k, \psi_{-i}(\theta_{-i})]$$

- Denote the truthful-reporting by ψ_i^*, i.e., $\psi_i^*(\theta_i) = \theta_i$ for each θ_i

Lemma 2.6. *The local ICs are sufficient for the global ICs.*

Proof. By Lemma 2.3 we know **p**-monotonicity is necessary. Consider the downard ICs when babbling is ψ_{-i}. By a similar argument in Lemma 2.3, for any $m > k > l$, if θ_i^k does not have incentive to report θ_i^l, then θ_i^m either. Thus, it is sufficient to see the case when $k = l+1$. □

By this lemma, from now on, we focus on the local upward and downward ICs. Consider agent i's local incentive compatibility involving θ^k and θ^{k-1} when the other employs a babbling $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$

$$\begin{aligned}
&p_i \left(\theta_i^k \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\
&\qquad + (1 - p_i) \left(\theta_i^k \Delta Q_i(\theta_i^k, \psi_{-i}) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \psi_i(\theta_{-i}) \right] \right) \\
&\geq p_i \Delta T_i(\theta_i^k, \psi_{-i}^*) + (1 - p_i) \Delta T_i(\theta_i^k, \psi_{-i}) \\
&\geq p_i \left(\theta_i^{k-1} \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\
&\qquad + (1 - p_i) \left(\theta_i^{k-1} \Delta Q_i(\theta_i^k, \psi_{-i}) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^{k-1}, \psi_i(\theta_{-i}) \right] \right)
\end{aligned}$$

Note that how the independent assumption simplifies this expression. In particular, when ψ_{-i} is a uniform babbling, i.e., $\psi_{-i}(\theta_{-i}) = \theta'_{-i}$ for all θ_{-i},

$$\begin{aligned} & p_i \left(\theta_i^k \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\ & \qquad + (1-p_i) \left(\theta_i^k \Delta q_i(\theta_i^k, \theta'_{-i}) + \gamma \left(\sum_{j \neq i} \mu_j \right) \Delta q_i(\theta_i^k, \theta'_{-i}) \right) \\ & \geq p_i \Delta T_i(\theta_i^k, \psi_{-i}^*) + (1-p_i) \Delta \tau_i(\theta_i^k, \theta'_{-i}) \\ & \geq p_i \left(\theta_i^{k-1} \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\ & \qquad + (1-p_i) \left(\theta_i^{k-1} \Delta q_i(\theta_i^k, \theta'_{-i}) + \gamma \left(\sum_{j \neq i} \mu_j \right) \Delta q_i(\theta_i^k, \theta'_{-i}) \right) \end{aligned}$$

First of all, given a babbling ψ_{-i}, to have τ_i to satisfy the inequality, the LHS should be (weakly) bigger than the RHS; this amounts to

$$p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i) \Delta Q_i(\theta_i^k, \psi_{-i})$$

is positive. As it should be the case for any agent, any different two types, this means that **p**-monotonicity is a necessary condition.

Note also that for any babbling of $-i$, the expected payment should be a linear combination of those of the uniform babblings, i.e.,

$$\begin{aligned} & p_i \Delta T_i(\theta_i^k, \psi_{-i}^*) + (1-p_i) \Delta T_i(\theta_i^k, \psi_{-i}) \\ & \qquad = \sum_{\theta_{-i}} \lambda(\theta_{-i}) (p_i \Delta T_i(\theta_i^k, \psi_{-i}^*) + (1-p_i) \Delta \tau_i(\theta_i^k, \psi_{-i}(\theta_{-i}))) \end{aligned} \quad (2.14)$$

Given a babbling ψ_{-i}, denote the right end of the interval which the corresponding inequality induces by $R(\psi_{-i})$ and the left end by $L(\psi_{-i})$; with an abuse of notation, for a uniform babbling $\psi_{-i}(\theta_{-i}) = \theta'_{-i}$ for all θ_{-i}; denote each by $R(\theta'_{-i})$ and $L(\theta'_{-i})$.

Our question is whether we can find $\tau_i(\theta_i^k, \theta'_{-i})$ and $\tau_i(\theta_i^{k-1}, \theta'_{-i})$ (or equivalently, $\Delta \tau_i(\theta_i^k, \theta'_{-i})$) for each θ'_{-i} which satisfies all the local incentive compatibilities (i.e., for every babbling of $-i$).

Lemma 2.7. *By the linear relationship (2.14), such τ_i exists if and only if for each ψ_{-i},*

$$[\bar{L}(\psi_{-i}), \bar{R}(\psi_{-i})] \cap [L(\psi_{-i}), R(\psi_{-i})] \neq \varnothing \tag{2.15}$$

where

$$\bar{L}(\psi_{-i}) \equiv \sum_{\theta_{-i}} \lambda(\theta_{-i}) L(\psi_{-i}(\theta_{-i}))$$
$$\bar{R}(\psi_{-i}) \equiv \sum_{\theta_{-i}} \lambda(\theta_{-i}) R(\psi_{-i}(\theta_{-i}))$$

i.e., each of them are a linear combination of the left (the right) end of the uniform babblings, corresponding to θ_{-i}.

Proof. The necessary part is obvious. For the sufficiency, we can first pin down each $\Delta\tau_i(\theta_i^k, \theta_{-i})$ for each θ_{-i} using the inequality for uniform babbling. By the condition, we know that this satisfies the other inequalities for non-uniform babblings. □

Note that condition (2.15) is satisfied if and only if

$$\min\{\bar{R}(\psi_{-i}), R(\psi_{-i})\} \geq \max\{\bar{L}(\psi_{-i}), L(\psi_{-i})\} \tag{2.16}$$

Suppose $\bar{R}(\psi_{-i}) \geq R(\psi_{-i})$. Note that this implies $\bar{L}(\psi_{-i}) \geq L(\psi_{-i})$. Then, (2.16) amounts to

$$\begin{aligned} & p_i \left(\theta_i^k \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\ & \qquad + (1 - p_i) \left(\theta_i^k \Delta Q_i(\theta_i^k, \psi_{-i}) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}) \right] \right) \\ & \geq p_i \left(\theta_i^{k-1} \Delta Q_i(\theta_i^k, \psi_{-i}^*) + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \Delta q_i(\theta_i^k, \theta_{-i}) \right] \right) \\ & \qquad + (1 - p_i) \left(\theta_i^{k-1} \Delta Q_i(\theta_i^k, \psi_{-i}) + \gamma \left(\sum_{j \neq i} \mu_j \right) \Delta Q_i(\theta_i^k, \psi_{-i}) \right) \end{aligned}$$

if and only if

$$\begin{aligned}&(\theta_i^k - \theta_i^{k-1})(p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i}))\\&\geq (1-p_i)\gamma \sum_{\theta_{-i}} \lambda(\theta_{-i}) \left(\sum_{j \neq i} (\theta_j - \mu_j)\right) \Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))\\&= (1-p_i)\gamma \mathbb{E}_{\theta_{-i}} \left[\sum_{j \neq i} (\theta_j - \mu_j) \Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))\right].\end{aligned}$$

Similarly, if we also consider the case $R(\psi_{-i}) \geq \bar{R}(\psi_{-i})$, altogether we have

$$\begin{aligned}&(\theta_i^k - \theta_i^{k-1})(p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i}))\\&\geq (1-p_i)\gamma \left|\mathbb{E}_{\theta_{-i}}[\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i})) \sum_{j \neq i} (\theta_j - \mu_j)]\right|.\end{aligned}$$

Rearranging this, and noting the condition should hold any ψ_{-i},

$$\theta_i^k - \theta_i^{k-1} \geq (1-p_i)\gamma \max_{\psi_{-i}} \left|\sum_{\theta_{-i}} \frac{\lambda(\theta_{-i})\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))}{p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \sum_{j \neq i} (\theta_j - \mu_j)\right|.$$

This completes the necessary part of the proof. Suppose q satisfies **p**-monotonicity and for each θ^k,

2.9.1.6 Proof of Corollary 2.2

Proof. Note that in this case $\Delta Q_i(\theta_i^k, \psi_{-i}) = \Delta q_i(\theta_i^k, \theta_{-i})$ for all ψ_{-i}. Then, for each ψ_{-i},

$$\begin{aligned}&\left|\sum_{\theta_{-i}} \frac{(1-p_i)\lambda(\theta_{-i})\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))}{p_i \Delta Q_i(\theta_i^k, \psi_{-i}^*) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \sum_{j \neq i} (\theta_j - \mu_j)\right|\\&= \left|(1-p_i) \sum_{\theta_{-i}} \lambda_i(\theta_{-i}) \sum_{j \neq i} (\theta_j - \mu_j)\right| = 0.\end{aligned}$$

□

2.9.1.7 Proof of Corollary 2.3

Proof. Note that for each ψ_{-i},

$$\begin{aligned}
&\left|\sum_{\theta_{-i}} \frac{(1-p_i)\lambda(\theta_{-i})\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))}{p_i\Delta Q_i(\theta_i^k, \psi^*_{-i}) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \sum_{j\neq i}(\theta_j - \mu_j)\right| \\
&\leq \sum_{\theta_{-i}} \frac{(1-p_i)\lambda(\theta_{-i})\Delta q_i(\theta_i^k, \psi_{-i}(\theta_{-i}))}{p_i\Delta Q_i(\theta_i^k, \psi^*_{-i}) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \left|\sum_{j\neq i}(\theta_j - \mu_j)\right| \\
&\leq \frac{(1-p_i)\lambda(\theta_{-i})\Delta Q_i(\theta_i^k, \psi_{-i})}{p_i\Delta Q_i(\theta_i^k, \psi^*_{-i}) + (1-p_i)\Delta Q_i(\theta_i^k, \psi_{-i})} \max_{\psi_{-i}} \left|\sum_{j\neq i}(\theta_j - \mu_j)\right| \\
&= \frac{1}{\frac{p_i}{1-p_i}\frac{\Delta Q_i(\theta_i^k, \psi^*_{-i})}{\Delta Q_i(\theta_i^k, \psi_{-i})} + 1} \max_{\psi_{-i}} \left|\sum_{j\neq i}(\theta_j - \mu_j)\right|
\end{aligned}$$

Lemma 2.8. *For any $q_i(\theta_i,)$,*

$$\frac{\Delta Q_i(\theta_i, \psi'_{-i})}{\Delta Q_i(\theta_i, \psi_{-i})} \geq \frac{\min_{\theta_{-i}} \lambda_i(\theta_{-i})}{\max_{\theta_{-i}} \lambda_i(\theta_{-i})}, \forall \psi_{-i}, \psi'_{-i}.$$

Also, the bound is tight.

Proof. For the proof of the lemma, we rename each θ_{-i} so that

$$\lambda_i(\theta^1_{-i}) \geq \lambda_i(\theta^2_{-i}) \geq \cdots \geq \lambda_i(\theta^M_{-i})$$

and

$$\Delta q_i(\theta_i, \theta_{-i,1}) \geq \Delta q_i(\theta_i, \theta_{-i,2}) \geq \cdots \geq \Delta q_i(\theta_i, \theta_{-i,M})$$

(so we reorder θ_{-i} in two different ways) where $M = \prod_{j\neq i} |\Theta_j|$. Note that to minimize the fraction we should minimize the numerator and maximize the denominator. To this end, the numerator should be "anti-assortative", while the denominator should be "assortative":

$$\frac{\lambda_i(\theta^1_{-i})\Delta q_i(\theta_i, \theta_{-i,M}) + \lambda_i(\theta^2_{-i})\Delta q_i(\theta_i, \theta_{-i,M-1}) + \cdots + \lambda_i(\theta^M_{-i})\Delta q_i(\theta_i, \theta_{-i,1})}{\lambda_i(\theta^1_{-i})\Delta q_i(\theta_i, \theta_{-i,1}) + \lambda_i(\theta^2_{-i})\Delta q_i(\theta_i, \theta_{-i,2}) + \cdots + \lambda_i(\theta^M_{-i})\Delta q_i(\theta_i, \theta_{-i,M})}$$

By the non-constant assumption, there exits i and θ_i such that $\Delta q_i(\theta_i, \theta_{-i}, 1) > 0$. Note that by the independent assumption, $\lambda_i(\theta^M_{-i}) > 0$. For notational convenience, rewrite the expression as

$$\frac{c_1 x_M + c_2 x_{M-1} + \cdots + c_M x_1}{c_1 x_1 + c_2 x_2 + \cdots + c_M x_M}$$

where $c_m = \lambda_i(\theta^m_{-i})$ and $x_m = \Delta q_i(\theta_i, \theta_{-i,m})$. Recall that $x_1 > 0$. We first show that $x_M = 0$. Note that

$$\frac{d}{dx_M}\left(\frac{c_1 x_M + y}{z + c_M x_M}\right) = \frac{c_1 z - c_M y}{(z + c_M x_M)^2}$$

where

$$y \equiv c_2 x_{M-1} + \cdots + c_M x_1$$

$$z \equiv c_1 x_1 + c_2 x_2 + \cdots + c_{M-1} x_{M-1}$$

Note that $z \geq y > 0$ (the last inequality comes from $x_1 > 0$) and thus if $c_1 > c_M$ then the equality is strict. Thus, the value of the derivative is positive. Note that $c_1 = c_M$ requires $c_1 = c_2 = \cdots = c_M$. Suppose $c_1 > c_M$; then $x_M = 0$ as the derivative is strictly positive. If $c_1 = c_m$ and $z = y$, then the derivative is 0; so in this case, our choice of c_M does not affect the value. So, we choose $x_M = 0$. Then, we have

$$\frac{c_2 x_{M-1} + \cdots + c_M x_1}{c_1 x_1 + \cdots + c_{M-1} x_{M-1}}$$

we apply the same argument to have $x_{M-1} = 0$. Repeat inductively this step until we have only x_1. As a result, we have

$$\frac{c_M x_1}{c_1 x_1} = \frac{c_M}{c_1}.$$

□

□

2.9.1.8 Proof of Lemma 2.4

Proof. We first show that an efficient allocation rule satisfies ex-post monotonicity if and only if $\gamma \geq 1$. First we show that "if" part: we claim that for each i, $q^*_i(\theta_i, \theta_{-i}) > 0$ only if $\theta_i \geq \max_{j \neq i} \theta_j$. For the contradiction, suppose not; there exists $j \neq i$ such that $\theta_j > \theta_i$. Then, it increases the social surplus to reduce $q^*_i(\theta_i, \theta_{-i})$ and increase $q^*_j(\theta_i, \theta_{-i})$ (which is strictly less than 1), because by doing so the net change is

$$\theta_j + \gamma\theta_i - \theta_i - \gamma\theta_j = (1 - \gamma)(\theta_j - \theta_i).$$

For a similar reason, if $\theta_i > \max_{j\neq i}\theta_j$, then $q_i^*(\theta_i, \theta_{-i}) = 1$.

Next, we prove "only if" part: Suppose $\gamma > 1$. By the non-constant qualification and ex-post monotonicity, there exists $i \in \mathcal{I}$ and θ_i', θ_i with $\theta_i' > \theta_i$ and θ_{-i} such that $q_i^*(\theta_i', \theta_{-i}) > q_i^*(\theta_i, \theta_{-i})$. Then, note that for some $j \neq i$, $\theta_j + \gamma\sum_{k\neq j}\theta_k \geq \theta_i + \gamma\sum_{j\neq i}\theta_j$. Otherwise, $q_i^*(\theta_i, \theta_{-i}) = 1$ for efficiency; a contradiction. From this and $\gamma > 1$, we know that

$$\theta_j + \gamma\theta_i \geq \theta_i + \gamma\theta_j \iff \theta_i \geq \theta_j$$

thus $\theta_i' > \theta_j$. This implies that

$$\theta_j + \gamma\theta_i' > \theta_i' + \gamma\theta_j \iff \theta_j + \gamma\sum_{k\neq j}\theta_k \geq \theta_i' + \gamma\sum_{j\neq i}\theta_j$$

Thus, $q_i^*(\theta_i', \theta_{-i}) = 0$. Contradiction. Given this we know that 0-monotonicity implies any $\mathbf{p}$-monotonicity. This completes the proof for the first part of the lemma. □

2.9.1.9 Proof of Proposition 2.4

Proof. Let us make the following simple observations:

Claim 2.1. $w_1 = 1/2$.

To see this, suppose $w_1 < 1/2$. Then, we can increase w_1 and increase the objective without sacrificing any conditions.

Claim 2.2. $x_1 + y_1 = 1$.

Suppose it is not true. Then, increase x_1, which increases the objective.

Claim 2.3. As long as $z_1 \leq 1/2$, which is true from (2.5), (2.4) implies (2.3).

To see this, note that if

$$(1-p)w_1 - (1-p)y_1 \leq (1-p)x_1 - (1-p)z_1$$

Then, (2.4) implies (2.3). As $w_1 = 1/2$ by Claim 1, and $x_1 = 1 - y_1$ by Claim 2,

$$\frac{1}{2} - y_1 \leq (1 - y_1) - z_1 \iff z_1 \leq \frac{1}{2}.$$

Claim 4: (2.4) should be hold with equality.

By Claim 3, we know that we can ignore (2.3). Then, note that when $y_1 = 1$ (so, $x_1 = 0$),

$$p\left(\frac{1}{2}0 + \frac{1}{2}\frac{1}{2}\right) + (1-p)\frac{1}{2} < p\left(\frac{1}{2}z_1 + \frac{1}{2}\right) + (1-p)$$

Thus, y_1 should be less than 1. Thus, if the inequality does not hold with equality, we could increase y_1 slightly so as to increase the objective without violating the inequality.

Thus,

$$\begin{aligned} & p\left(\frac{1}{2}(1-y_1) + \frac{1}{4}\right) + (1-p)\frac{1}{2} = p\left(\frac{1}{2}z_1 + \frac{1}{2}y_1\right) + (1-p)y_1 \\ \iff\ & y_1(-p-(1-p)) = -p\left(\frac{1}{2} + \frac{1}{4}\right) - (1-p)\frac{1}{2} + p\frac{1}{2}z_1 \\ \iff\ & y_1 = p\frac{3}{4} + (1-p)\frac{1}{2} - p\frac{1}{2}z_1. \end{aligned}$$

Substituting it into the objective,

$$\begin{aligned} \frac{1}{2}\left(3z_1 + 31y_1 + 17(1-y_1) + \frac{45}{2}\right) &= \frac{1}{2}\left(3z_1 + 14\left(p\frac{3}{4} + (1-p)\frac{1}{2} - p\frac{1}{2}z_1\right) + 17 + \frac{45}{2}\right) \\ &= \frac{1}{2}\left(z_1\left(3 - \frac{14}{2}p\right) + 14\left(p\frac{3}{4} + (1-p)\frac{1}{2}\right) + 17 + \frac{45}{2}\right). \end{aligned}$$

As it is linear in z_1, if $\left(3 - 4 - \frac{14}{2}p\right) \geq 0$ or $p \leq \frac{3}{7}$ then $z_1 = \frac{1}{2}$, while $z_1 = 0$ when $p > \frac{3}{7}$. □

2.9.1.10 Proof of Proposition 2.5

Proof. Consider agent i's first order condition:

$$p_i \mathbb{E}_{\theta_{-i}}\left[\frac{u_i(f(\theta_i,\theta_{-i}),\theta_i)}{\partial\theta_i}\right] + (1-p_i)\frac{\partial u_i(f(\theta_i,\theta'_{-i}),\theta_i)}{\partial\theta_i} = 0, \forall\theta'_{-i}$$

Suppose for some θ''_{-i},

$$\frac{\partial u_i(f(\theta_i,\theta''_{-i}),\theta_i)}{\partial\theta_i} = k \neq 0.$$

From the FOC, it implies that for any θ'_{-i}, $\frac{\partial u_i(f(\theta_i,\theta''_{-i}),\theta_i)}{\partial\theta_i} = k$. In turn, this also implies $\mathbb{E}_{\theta_{-i}}\left[\frac{u_i(f(\theta_i,\theta_{-i}),\theta_i)}{\partial\theta_i}\right] = k$; thus the LHS of the FOC amounts to $k \neq 0$; a contradiction. □

2.9.1.11 Proof of Proposition 2.6

Proof. Consider agent i's first order condition:

$$p_i\mathbb{E}_{\theta_{-i}}\left[\frac{u_i(f(\theta_i,\theta_{-i}),\theta_i,\theta_{-i})}{\partial\theta_i}\right]+(1-p_i)\mathbb{E}_{\theta_{-i}}\left[\frac{\partial u_i(f(\theta_i,\psi_{-i}(\theta_{-i})),\theta_i,\theta_{-i})}{\partial\theta_i}\right]=0,\forall\psi_{-i}:\Theta_{-i}\to\Theta_{-i}$$

Suppose for some $\psi'_{-i}:\Theta_{-i}\to\Theta_{-i}$,

$$\mathbb{E}_{\theta_{-i}}\left[\frac{\partial u_i(f(\theta_i,\psi'_{-i}(\theta_{-i})),\theta_i)}{\partial\theta_i}\right]=k\neq 0.$$

From the FOC, it implies that for any ψ_{-i}, $\mathbb{E}_{\theta_{-i}}\left[\frac{\partial u_i(f(\theta_i,\psi_{-i}(\theta_{-i})),\theta_i)}{\partial\theta_i}\right]=k$. In particular, when $\psi_{-i}(\theta_{-i})=\theta_{-i}$ (i.e., babbling reports true type), i.e.,

$$\mathbb{E}_{\theta_{-i}}\left[\frac{u_i(f(\theta_i,\theta_{-i}),\theta_i,\theta_{-i})}{\partial\theta_i}\right]=k$$

This implies LHS of the FOC is equal to $k\neq 0$; a contradiction. □

2.9.1.12 Proof of Proposition 2.7

Proof. Because q_i satisfies slope independence, we may write it as

$$q_i(\theta_i,\theta_{-i})=k_i(\theta_{-i})+y_i(\theta_i)$$

for some $k_i:\Theta_{-i}\to[0,1]$ and $y_i:\Theta_i\to[0,1]$ s.t. $k_i(\theta_{-i})+y_i(\theta_i)\in[0,1]$ for all $\theta_{-i}\in[0,1]^{N-1}$ and $\theta_i\in[0,1]$ and y_i is increasing. Fix $\theta'_{-i}\in\Theta_{-i}$,

$$\begin{aligned}U_i(\theta_i;\theta'_{-i})&\equiv\mathbb{E}_{\theta_{-i}}\left[(\theta_i+\gamma\theta_{-i})(k_i(\theta'_{-i})+y_i(\theta_i))\right]-\tau_i(\theta_i,\theta'_{-i})\\&=\left(\theta_i+\gamma\left(\sum_{j\neq i}\mu_j\right)\right)(k_i(\theta'_{-i})+y_i(\theta_i))-\tau_i(\theta_i,\theta'_{-i})\end{aligned}$$

From this, by the standard argument, we can show that U_i is concave in θ_i and thus it is almost everywhere differentiable; and

$$U'_i(\theta_i,\theta'_{-i})=k_i(\theta'_{-i})+y_i(\theta_i)$$

From this

$$U_i(\theta_i;\theta'_{-i})=U_i(0,\theta'_{-i})+k_i(\theta'_{-i})\theta_i+\int_0^{\theta_i}y_i(x)dx$$

where $U_i(0,\theta'_{-i}) \in \mathbb{R}$. Define

$$\tau_i(\theta_i,\theta'_{-i}) := \left(\theta_i + \gamma\left(\sum_{j\neq i}\mu_j\right)\right)(k_i(\theta'_{-i}) + y_i(\theta_i)) - U_i(0,\theta'_{-i}) - k_i(\theta'_{-i})\theta_i - \int_0^{\theta_i} y_i(x)dx.$$

We shall show that with this transfer rule, agent i's incentive compatibility holds for any $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$. Fix $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$. Let us consider the relevant incentive compatibility of agent i between reporting θ_i and θ'_i (note that we are considering global incentive constraints)

$$\begin{aligned}
&\mathbb{E}_{\theta_{-i}}\left[\left(\theta_i+\gamma\sum_{j\neq i}\theta_j\right)(k_i(\psi_{-i}(\theta_{-i})+y_i(\theta_i))\right] - \mathbb{E}_{\theta_{-i}}[\tau_i(\theta_i,\psi_{-i}(\theta_{-i}))] \\
&\qquad \geq \mathbb{E}_{\theta_{-i}}\left[\left(\theta_i+\gamma\sum_{j\neq i}\theta_j\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta'_i))\right] - \mathbb{E}_{\theta_{-i}}[\tau_i(\theta'_i,\psi_{-i}(\theta_{-i}))]
\end{aligned}$$

$$\begin{aligned}
\iff &\mathbb{E}_{\theta_{-i}}\left[\left(\theta_i+\gamma\sum_{j\neq i}\theta_j\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta_i))\right] - \mathbb{E}_{\theta_{-i}}[(\theta_i+\gamma\sum_{j\neq i}\mu_j)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta_i))] \\
&\qquad + U_i(0,\psi_{-i}) + k_i(\psi_{-i})\theta_i + \int_0^{\theta_i} y_i(x)dx \\
\geq\ &\mathbb{E}_{\theta_{-i}}\left[\left(\theta_i+\gamma\sum_{j\neq i}\mu_j\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta'_i))\right] - \mathbb{E}_{\theta_{-i}}\left[\left(\theta'_i+\gamma\sum_{j\neq i}\mu_j\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta'_i))\right] \\
&\qquad + U_i(0,\psi_{-i}) + K_i(\psi_{-i})\theta'_i + \int_0^{\theta'_i} y_i(x)dx
\end{aligned}$$

$$\begin{aligned}
\iff &\mathbb{E}_{\theta_{-i}}\left[\left(\sum_{j\neq i}\theta_j-\sum_{j\neq i}\mu_j\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta_i)\right] + U_i(0,\psi_{-i}) + k_i(\psi_{-i})\theta_i + \int_0^{\theta_i} y_i(x)dx \\
&\geq \theta_i k_i(\psi_{-i}) + \theta_i y_i(\theta'_i) + \mathbb{E}_{\theta_{-i}}\left[\gamma\left(\sum_{j\neq i}\theta_j\right)k_i(\psi_{-i}(\theta_{-i}))\right] + \gamma(\sum_{j\neq i}\mu_j)y_i(\theta'_i) \\
&\quad - \left(\theta'_i k_i(\psi_{-i}) + \theta'_i y_i(\theta'_i) + \gamma(\sum_{j\neq i}\mu_j)\mathbb{E}_{\theta_{-i}}[k_i(\psi_{-i}(\theta_{-i}))] + \gamma\left(\sum_{j\neq i}\mu_j\right)y_i(\theta'_i)\right) \\
&\qquad + U_i(0,\psi_{-i}) + K_i(\psi_{-i})\theta'_i + \int_0^{\theta'_i} y_i(x)dx
\end{aligned}$$

$$\begin{aligned}
\iff &\mathbb{E}_{\theta_{-i}}\left[\left(\sum_{j\neq i}(\theta_j-\mu_j)\right)(k_i(\psi_{-i}(\theta_{-i}))+y_i(\theta_i)\right] + U_i(0,\psi_{-i}) + k_i(\psi_{-i})\theta_i + \int_0^{\theta_i} y_i(x)dx \\
\geq\ &(\theta_i-\theta'_i)K_i(\psi_{-i}) + (\theta_i-\theta'_i)y_i(\theta'_i) + \mathbb{E}_{\theta_{-i}}\left[\gamma\sum_{j\neq i}(\theta_j-\mu_j)k_i(\psi_{-i}(\theta_{-i}))\right] + U_i(0,\psi_{-i}) + k_i(\psi_{-i})\theta'_i + \int_0^{\theta'_i} y_i(x)dx
\end{aligned}$$

$$\iff \int_0^{\theta_i} y_i(x)dx \geq (\theta_i - \theta_i')y_i(\theta_i') + \int_0^{\theta_i'} y_i(x)dx$$

$$\iff \int_{\theta_i}^{\theta_i'} y_i(x)dx \leq (\theta_i' - \theta_i)y_i(\theta_i')$$

where

$$U_i(0, \psi_{-i}) \equiv \mathbb{E}_{\theta_{-i}}[U_i(0, \psi_{-i}(\theta_{-i}))]$$

$$K_i(\psi_{-i}) \equiv \mathbb{E}_{\theta_{-i}}[k_i(\psi_{-i}(\theta_{-i})].$$

This holds as y_i is increasing.

□

2.9.1.13 Proof of Proposition 2.8

Proof. Suppose not, i.e., there exists $i \in \mathcal{I}$, θ_i and $\psi_{-i} : \Theta_{-i} \to \Theta_{-i}$ such that

$$\gamma \mathbb{E}_{\theta_{-i}} \left[\mu \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} \right] \neq \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} \right] \tag{2.17}$$

Suppose there is a transfer rule $\tau = (\tau_i)_i$ together which q is a **0**-dominant mechanism. Consider the first order condition for i, θ_i:

$$\mathbb{E}_{\theta_{-i}} \left[\theta_i \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} \right] + \gamma \mathbb{E}_{\theta_{-i}} \left[\left(\sum_{j \neq i} \theta_j \right) \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} \right] \\ - \mathbb{E}_{\theta_{-i}} \left[\frac{\partial \tau_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} \right] = 0. \tag{2.18}$$

Consider also the uniform babbling $\psi_{-i}(\cdot) = \theta_{-i}'$ for some $\theta_{-i}' \in [0,1]^{N-1}$. In this case, the first order condition is

$$\theta_i \frac{\partial q_i(\theta_i', \theta_{-i}')}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} + \gamma \sum_{j \neq i} \mu_j \frac{\partial q_i(\theta_i', \theta_{-i}')}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} - \frac{\partial \tau_i(\theta_i', \theta_{-i}')}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} = 0.$$

We can obtain a similar equation for any $\theta_{-i}' \in [0,1]^{N-1}$. In particular, for a given θ_{-i},

$$\theta_i \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} + \gamma \mu \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} - \frac{\partial \tau_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'} \bigg|_{\theta_i'=\theta_i} = 0$$

And so,

$$\int_{[0,1]^{N-1}} \theta_i \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i} \lambda_i(\theta_{-i})d\theta_{-i} + \int_{[0,1]^{N-1}} \gamma \sum_{j\neq i} \mu_j \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i} \lambda_i(\theta_{-i})d\theta_{-i}$$

$$- \int_{[0,1]^{N-1}} \frac{\partial \tau_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i} \lambda_i(\theta_{-i})d\theta_{-i} = 0$$

$$\iff \mathbb{E}_{\theta_{-i}}\left[\theta_i \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i}\right] + \gamma \mathbb{E}_{\theta_{-i}}\left[\sum_{j\neq i} \mu_j \frac{\partial q_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i} \lambda_i(\theta_{-i})\right]$$

$$-\mathbb{E}_{\theta_{-i}}\left[\frac{\partial \tau_i(\theta_i', \psi_{-i}(\theta_{-i}))}{\partial \theta_i'}\bigg|_{\theta_i'=\theta_i}\right] = 0.$$

Comparing this equation with (2.18) leads to a contradiction to (2.17). □

2.9.1.14 Proof of Lemma 2.5

Proof. Suppose condition (2.6) holds. Under our assumption $\gamma > 0$, it can be rewritten

$$\sup_{i\in\mathcal{I},\theta_i,\psi_{-i}} \left| \mathbb{E}_{\theta_{-i}}\left[\sum_{j\neq i}(\theta_j - \mu_j)\frac{\partial q_i(\theta_i, \psi_{-i}(\theta_{-i}))}{\partial \theta_i}\right]\right| = 0 \tag{2.19}$$

Pick $j \in \mathcal{I} \setminus \{i\}$ and $z \in [0,1]$ and define ψ_{-i} as follows:

$$\psi_{-i}(\theta_{-i}) := \begin{cases} \theta_{-i}' & \text{if } \theta_j \leq z, \\ \theta_{-i}'' & \text{if o.w.} \end{cases}$$

Then,

$$\begin{aligned}
&\mathbb{E}_{\theta_{-i}}\left[\sum_{j\neq i}(\theta_j - \mu_j)\frac{\partial q_i(\theta_i, \psi_{-i}(\theta_{-i}))}{\partial \theta_i}\right] \\
&= \int_{\theta_j\in[0,z]} \int_{\theta_k\in[0,1],\forall k\neq i,j} \sum_{l\neq i}(\theta_l - \mu_l)\frac{\partial q_i(\theta_i, \theta_{-i}')}{\partial \theta_i} \prod_{l\neq i} \lambda_l(\theta_l)d\theta_{-i} \\
&\quad + \int_{\theta_j\in[z,1]} \int_{\theta_k\in[0,1],\forall k\neq i,j} \sum_{l\neq i}(\theta_l - \mu_l)\frac{\partial q_i(\theta_i, \theta_{-i}'')}{\partial \theta_i} \prod_{l\neq i} \lambda_l(\theta_l)d\theta_{-i} \\
&= \int_{\theta_j\in[0,z]} (\theta_j - \mu_j)\frac{\partial q_i(\theta_i, \theta_{-i}')}{\partial \theta_i}\lambda_i(\theta_j)d\theta_j + \int_{\theta_j\in[z,1]} (\theta_j - \mu_j)\frac{\partial q_i(\theta_i, \theta_{-i}'')}{\partial \theta_i}\lambda_i(\theta_j)d\theta_j \\
&= \frac{\partial q_i(\theta_i, \theta_{-i}')}{\partial \theta_i}\int_0^1 (\theta_j - \mu_j)\lambda(\theta_j)d\theta_j + \left(\frac{\partial q_i(\theta_i, \theta_{-i}'')}{\partial \theta_i} - \frac{\partial q_i(\theta_i, \theta_{-i}')}{\partial \theta_i}\right)\int_z^1 (\theta_j - \mu_j)\lambda_i(\theta_j)d\theta_j \\
&= \left(\frac{\partial q_i(\theta_i, \theta_{-i}'')}{\partial \theta_i} - \frac{\partial q_i(\theta_i, \theta_{-i}')}{\partial \theta_i}\right)\int_z^1 (\theta_j - \mu_j)\lambda_i(\theta_j)d\theta_j
\end{aligned}$$

where the second equality comes from the fact for any $k \neq i, j$, $\int_0^1 \theta_k \lambda_k(\theta_k) d\theta_k = \mu_k$. The value of the last expression should be 0 for *any* $z \in [0, 1]$ in order to satisfy (2.19), which is impossible. □

2.9.1.15 Proof of Proposition 2.10

We first make two observations to prove the proposition.

Definition 2.12 (no waste)**.** An allocation rule $q = (q_i)_i$ satisfies *no waste* if

$$\sum_{i \in \mathcal{I}} q_i(\theta) = 1, \forall \theta \in \Theta.$$

Lemma 2.9. *Assume $N = 2$. An allocation rule $q = (q_i)_{i \in \mathcal{I}}$ satisfies no waste, slope independence, and monotonicity if and only if there exists $(y_i(\theta_i))_{i \in \mathcal{I}}$ where $y_i : \Theta_i \to \mathbb{R}$ such that*

(1) Additive separability: for each $i \in \mathcal{I}$,

$$q_i(\theta_i, \theta_j) = c_i + y_i(\theta_i) - y_j(\theta_j), j \neq i$$

where $c_1 + c_2 = 1$

(2) Monotonicity: for each agent i, $y_i(\cdot)$ is increasing in θ_i

(3) Feasibility: for each $i \in \mathcal{I}$ and $j \neq i$,

$$q_i(1, 0) = c_i + y_i(1) - y_j(0) \leq 1 \qquad (2.20)$$

$$q_i(0, 1) = c_i + y_i(0) - y_j(1) \geq 0. \qquad (2.21)$$

Proof. By slope independence,

$$q_1(\theta_1, \theta_2) = k_1(\theta_2) + y_1(\theta_1)$$

for some $k_1 : \Theta_2 \to \mathbb{R}$ and $y_1 : \Theta_1 \to \mathbb{R}$ s.t. $k_1(\theta_2) + y_1(\theta_1) \in [0,1]$ for all $(\theta_1, \theta_2) \in [0,1]^2$. By the monotonicity, y_1 is clearly increasing. Without loss, let $k_1(\theta_2) := c_1 - y_2(\theta_2)$. Then, by the no waste condition,

$$\begin{aligned} q_2(\theta_1, \theta_2) &= 1 - (c_1 + y_1(\theta_1) - y_2(\theta_2)) \\ &= 1 - c_1 - y_1(\theta_1) - y_2(\theta_2). \end{aligned}$$

Let $c_2 \equiv 1 - c_1$. By the monotonicity, clearly y_2 is increasing. Then, from the monotonicity,

$$c_i + y_i(\theta_i) - y_j(\theta_j) \leq c_i + y_i(1) - y_j(0) \leq 1$$

and also

$$c_i + y_i(\theta_i) - y_j(\theta_j) \geq c_i + y_i(0) - y_j(1) \geq 0.$$

□

Lemma 2.10. *Assume $N = 2$. Suppose an allocation rule $q = (q_i)_i$ satisfies the additive separability, monotonicity and feasibility if and only if the following class of allocation rules that are characterized by $(k_i, \alpha_i, z_i)_i$ where $\alpha_i \in [0,1]$, $k_i \in [0,1]$ and $z_i : \Theta_i \to [0,1]$, which is increasing such that*

$$k_i + \sum_{i \in \mathcal{I}} \alpha_i = c_i + y_i(1) - y_i(0)$$

and

$$q_i(\theta_i, \theta_j) = k_i + \alpha_i z_i(\theta_i) + \alpha_j(1 - z_j(\theta_j)), \forall i, j \neq i.$$

They could be interpreted as follows: k_i is the probability that the item is always given to agent i. With probability α_i, the designer approaches to agent i and gives the item with probability $z_i(\theta) \in [0,1]$ to i; and gives it to j with the rest probability.

Proof. Consider an arbitrary allocation rule that satisfies the additive separability, monotonicity and feasibility (see Lemma 2.9), $q_1(\theta_1, \theta_2) = c_1 + y_1(\theta_1) - y_2(\theta_2)$ and $q_2(\theta_1, \theta_2) = c_2 + y_2(\theta_2) - y_1(\theta_1)$ where $c_1 + c_2 = 1$.

Conisider the case in which $y_i(1) > y_i(0)$ for all $i = 1, 2$. Then, $q_i(\theta)$ can be rewritten as

$$\begin{aligned} q_i(\theta_1, \theta_2) &= c_i + y_i(0) + (y_i(1) - y_i(0))\frac{y_i(\theta_i) - y_i(0)}{y_i(1) - y_i(0)} - y_j(1) + (y_j(1) - y_j(0))\frac{y_j(1) - y_j(\theta_j)}{y_j(1) - y_j(0)} \\ &= c_i + y_i(0) - y_j(1) + (y_i(1) - y_i(0))\frac{y_i(\theta_i) - y_i(0)}{y_i(1) - y_i(0)} + (y_j(1) - y_j(0))\left(1 - \frac{y_j(\theta_j) - y_j(0)}{y_j(1) - y_j(0)}\right). \end{aligned}$$

Define

$$\begin{aligned} k_i &\equiv c_i + y_i(0) - y_j(1) \\ \alpha_i &\equiv y_i(1) - y_i(0) \\ \alpha_j &\equiv y_j(1) - y_j(0) \\ z_i(\theta_i) &\equiv \frac{y_i(\theta_i) - y_i(0)}{y_i(1) - y_i(0)} \\ z_j(\theta_j) &\equiv \frac{y_j(\theta_j) - y_j(0)}{y_j(1) - y_j(0)}. \end{aligned}$$

For the case in which $y_i(0) = y_i(1) = y_i$, we choose $\alpha_i = 0$ and $k_i = c_i + y_i$. From (2.20) and (2.21):

$$\begin{aligned} c_i + y_i(1) - y_j(0) &\leq 1 \\ -c_i - y_i(0) + y_j(1) &\leq 0 \end{aligned}$$

by adding these two, we have $y_i(1) - y_i(0) - y_j(0) + y_j(1) \leq 1$; and the monotonicity of y_j implies $y_i(1) - y_i(0) \leq 1$.

Also, by monotonicity, z_i and z_j are in the range $[0, 1]$ and also increasing. With these definitions,

$$q_i(\theta_i, \theta_j) = k_i + \alpha_i z_i(\theta_i) + \alpha_j z_j(\theta_j).$$

Note that

$$\begin{aligned} k_i + \alpha_i + \alpha_j &= c_i + y_i(0) - y_j(1) + y_i(1) - y_i(0) + y_j(1) - y_j(0) \\ &= c_i + y_i(1) - y_j(0). \end{aligned}$$

That is, the maximum possible allocation for agent i given the monotonicity. □

To maximize social surplus, trivially q satisfies no waste condition. By Lemma 2.9 and Lemma 2.10, it is without loss to consider the following class of allocation rules:

$$q_i(\theta_i, \theta_j) = k_i + \alpha_i z_i(\theta_i) + \alpha_j(1 - z_j(\theta_j))$$

where $k_i \in [0,1]$, $\alpha_i \in [0,1]$, $k_i + k_j + \alpha_i + \alpha_j = 1$, $z_i : \Theta_i \to [0,1]$ and increasing. As mentioned, we interpret $z_i(\theta_i)$ is the probability of giving the item to agent i conditional on the designer having approached to agent i; and conditional on it, we know what is the most efficient allocation rule, which is characterized in Proposition 2.10; we can apply this result to each agent i and we know the resulting expected social surplus is V_i in (2.8). This means given we already choose k_i and α_i for each i, the maximum social surplus attainable is

$$k_i(\mu_i + \gamma\mu_j) + k_j(\mu_j + \gamma\mu_i) + \alpha_i V^i + \alpha_j V^j \tag{2.22}$$

By definition of V^i and V^j, $\min\{V^i, V^j\} \geq \max\{\mu_i + \gamma\mu_j, \mu_j + \gamma\mu_i\}$. In addition, the linearity of (2.22), we know the optimal social surplus is

$$V^* = \max_{i \in \mathcal{I}} V_i.$$

CHAPTER 3

Reputation and Information Disclosure with Bounded Memory

3.1 Introduction

The decision of information disclosure plays an important role in managing reputation.

In this paper, we consider a repeated game in which there is a long-run economic agent (e.g., a firm) who chooses the quality of its product each period *and* also decides whether to disclose the quality. There is a stream of short-run players (e.g., consumers), who choose between buying high or low volume of the product. The short-run player prefers buying high volume only when the quality of the product is expected to be high. Since the long-run player is assumed to be better to make low quality product regardless of the volume purchased, the static Nash equilibrium predicts low quality product purchased in a small volume, which is assumed to be payoff-dominated by the outcome of high quality and high volume.

In reality, a reputation mechanism often plays a role in this situation. We capture reputation concern of this agent by introducing incomplete information of its type Kreps et al. (1982); Fudenberg and Levine (1989): with some probability, there is a chance that the firm is the type which is "committed" to make a high quality product. Moreover, the commitment type discloses the quality with a certain probability. Given this, to build reputation, the long-run player should behave similar to the commitment type; especially, given that the short-run players only can observe the message sent by the long-run player, the long-run player should send messages similarly to the commitment type. What is important is to look "similar" to the commitment type: always making high quality product and disclosing it may degrade its reputation if the commitment type often does not disclose the quality.

In addition, for most parts of this paper, we assume that the short-run players only can observe a finite number of past histories. This assumption implies any past history can be "forgotten" by short-run players as time passes; and we expect that the long-run player's play would be substantially different based on the length of the observable history.

We first show that there is an equilibrium in which the long-run player's information disclosure is "fully informative": a firm having produced a high quality product discloses it, while the low quality product does not. This result may remind the reader of the well-known "unraveling" result by, for example, Milgrom (1981). However, in our case, the driving force is the long-run players' reputation concern, i.e., to make the short-run players believe it more likely as the commitment type, which could be exploited at some point in the future.

Focusing on such fully informative equilibria, we then characterize equilibrium dynamics. The equilibria are cyclical: the long-run player builds reputation and milk it. The detailed dynamics, including how often they can exploit and build reputation, is determined by the prior, the length of observable histories, and the information disclosure behavior of the commitment type.

The present paper is related to the information transmission literature. In particular, as mentioned, the fully informative equilibrium is similar to the result in the models with verifiable information (e.g., Grossman (1981), Milgrom (1981) and Okuno-Fujiwara, Postlewaite, and Suzumura (1990) among others). They commonly study the information disclosure scheme where the informed party cannot lie in the sense that the inverse of a message should contain the true state of nature. In some following results, we have a similar argument to their *unraveling* result, although ours is dynamic and reputational. Recently, Van Der Schaar and Zhang (2015) and Marinovic et al. (2018) study a relevant question to ours.[1]

The present paper is closely related to the reputation literature with bounded memory.

[1] In both papers, the quality is private information; the reputation means the posterior belief for the quality. In the former, the quality is given and fixed, while in the latter whose model is based on Board and Meyer-ter-Vehn (2013), the quality of the firm is persistent.

This assumption is employed by Liu (2011) and Ekmekci (2011). Both papers share the feature that short-run players obtain restricted information about the past play. In Liu (2011), each short-run player decides how many past periods he/she will observe, which is costly. Thus, the informational decision is made by the uninformed side. On the other hand, we are interested in information disclosure problem by the informed player. Ekmekci (2011) shows that reputation is sustainable by introducing an elaborately designed rating system which effectively incentivize long-run agent to play a targeted action (e.g., the Stackelberg action). These study, including ours, are related to the literature of explaining permanent reputation.[2] There has been studies under which environments a cyclic or permanent reputation is maintained. More recently, Bhaskar and Thomas (2018) studies a version of repeated games with bounded memory in the context of community enforcement. In recent papers, Jullien and Park (2014) and Mathevet et al. (2019) study dynamic information transmission for building reputation; unlike ours, they study "cheap talk" and do not assume bounded memory.

The rest of this paper is organized as follows: In Section 3.2 we set up the model which includes descriptions of the information disclosure regime which we call *binary disclosure*. Throughout Section 3.3, we make an analysis, presenting formal expositions of the results. Finally, in Section 3.5 we discuss a few possible extensions and generalization of this study. We delegate all the proofs to Appendix unless specified.

[2] The result of impermanent reputation is established by Cripps, Mailath, and Samuelson (2004), which shows that with imperfect monitoring, a short-run player almost surely learn the type of the long-run player, and the continuation play of any equilibrium is asymptotically Nash equilibrium of game with complete information. Fudenberg and Levine (1992) shows that the reputation effect gives the equilibrium payoff of long-run player is arbitrarily close to the Stackelberg payoff. This is not contradictory to Cripps, Mailath, and Samuelson (2004), since the payoff is evaluated *ex-ante* in the former.

3.2 The Model

3.2.1 The Stage Game

There are a long-run player and a stream of short-run players. We call them player 1 ("he") and 2 ("she"), respectively. In each period $t \in \{0, 1, 2, \dots\}$, player 1 and 2 simultaneously choose their (possibly mixed) action: player 1 decides whether to make high (H) quality or low (L) quality, i.e., $A_1 := \{H, L\}$, and player 2 chooses from $A_2 := \{h, l\}$ which could be interpreted as how much they purchase the product. This interpretation requires the product to be an experience good. After the realization of action, player 1 decides whether to disclose it (details follow).

We make the following assumptions on the stage game:

Assumption 3.1 (product-choice game). *The stage game has following payoff structure:*

(a) For all $a_2 \in A_2$, $u_1(L, a_2) > u_1(H, a_2)$.

(b) $u_1(L, h) - u_1(H, h) > u_1(L, l) - u_1(H, l)$

(c) $u_2(L, l) > u_2(L, h)$

(d) $u_2(H, h) > u_2(H, l)$

(e) $u_1(H, h) > u_1(L, l)$

Condition (a) says that for player 1, playing L is a dominant strategy in the stage game; (b) says that the opportunistic behavior, L, is more profitable when player 2 plays h with higher probability; Due to (c) and (a), the stage game Nash equilibrium is (L, l); With (c), (d) implies that there exists a critical value, $\bar{\alpha} \in (0, 1)$, of the probability of playing H by player 1 with this playing h and l is indifferent to player 2.[3] (e) captures that building reputation is beneficial. We call a stage game satisfying these assumptions a *product-choice game*. Figure 3.1 is an example of such games.

[3] There is a difference from this game with the Prisoners' Dilemma, since playing l is not a dominant strategy for player 2: if player 1 plays H with high enough probability, then it is better for player 2 to play h.

	h	l
H	$2, 3$	$0, 2$
L	$4, 0$	$1, 1$

Figure 3.1: An example of the product-choice game

Lemma 3.1. *If $\alpha_2, \alpha_2' \in \Delta(A_2)$ satisfies $\alpha_2[h] > \alpha_2'[h]$,*

$$u_1(a_1, \alpha_2) > u_1(a_1, \alpha_2'), \forall a_1 \in A_1.$$

Proof. From (a) and (d), we obtain

$$u_1(L, h) > u_1(H, h) > u_1(L, l) > u_1(H, l).$$

Then the conclusion immediately follows. □

3.2.2 Incomplete Information

There are 2 *types* of long-run player, the *normal* type and the *commitment* type $\Theta = \{\theta_n, \theta_c\}$. The normal type chooses his action and whether to disclose to maximize the lifetime expected payoff, where the stage game satisfies Assumption 3.1. On the other hand, the commitment type plays H in each period, while we assume that it randomly discloses its action with probability $\beta \in [0, 1]$. The players share the common prior, $\mu_0 \in \Delta(\Theta)$; with an abuse of notation, let $\mu_0 \equiv \mu_0(\theta_c)$.

3.2.3 Information Disclosure

In each period after the action is realized, player 1 chooses whether to publicly disclose truthfully or withhold it. By this, we implicitly assume that the realized action, which we interpret as the quality of the product, is verifiable (i.e., hard information). The message space is

$$M \equiv \{H, L, B\}$$

where B represents "blank." This assumption rules out the possibility of sending message H after the realized action is L, while allowing withhold this information. We shall formally introduce this assumption momentarily (Assumption 3.2). For example, player 1 cannot send message H when the realized action is L, while being allowed not to disclose it. This assumption is natural in many cases.[4] Importantly, we further assume the length of public history is exogenously given as $K \in \mathbb{N}$. This may reflect the cost of holding information or the government regulation.

3.2.4 Solution Concept

Let $\mathcal{H}_0 = \{\varnothing\}$,

$$\mathcal{H}_t \equiv (A_1 \times A_2 \times M)^t, t \geq 1 \text{ and } \mathcal{H} \equiv \bigcup_{t=0}^{\infty} \mathcal{H}_t.$$

We are interested in stationary public equilibria: Player 2's (stationary) strategy is $\sigma_2 : M^K \to \Delta(A_2)$. That is, we only consider strategies that do not depend on the calendar time of the game. Player 1's (behavioral) strategy is a function $\sigma_1 : \mathcal{H} \to \Delta(A_1)$. Along with this, there is additional information disclosure decision by player 1, $\sigma_{1m} : \mathcal{H} \times A_1 \to \Delta(M)$ where $M = \{H, L, B\}$. Because player 2's strategy only depends on M^K, we can restrict to a smaller set of strategies $\sigma_1 : M^K \to \Delta(A_1)$, and $\sigma_{1m} : M^K \times A_1 \to \Delta(M)$.

Assumption 3.2. *$M = \{H, L, B\}$ and $\sigma_{1m}(\cdot, a_1) \in \Delta\{a_1, B\}$ for all $a_1 \in A_1$.*

Player 1's maximization problem is

$$\max_{\sigma_1} \mathbb{E}_{\mathbf{P}} \left[(1-\delta) \sum_{t=0}^{\infty} \delta^t u_1(\sigma_1(a_1^t), \sigma_2(a_2^t)) \right]$$

where $\mathbf{P}$ is the probability measure induced by σ_1, σ_{1m} and σ_2 on $\mathcal{H}$, and $\delta \in [0, 1)$ is the discount factor of player 1.

Let $\mathcal{S} \equiv M^K$ be the set of *states*, and $p_0 \in \Delta(\mathcal{S})$ (res. $p_1 \in \Delta(\mathcal{S})$) be an invariant distribution of the normal type over $\mathcal{S}$ (res. the commitment type). Also let $\mu : \mathcal{S} \to [0, 1]$

[4] Similar assumption has been employed in many papers; for example, Grossman (1981) and Milgrom (1981) among others.

be the posterior belief for the commitment type at s, $\sigma_1 : \mathcal{S} \to \Delta(A_1)$, $\sigma_{1m} : \mathcal{S} \times A_1 \to \Delta(M)$ and $\sigma_2 : \mathcal{S} \to \Delta(A_2)$ be strategies which depend on the state. A $(\sigma_1, \sigma_{1m}, \sigma_2, \mu, p_0)$ is a *stationary perfect Bayesian equilibrium* if

1. Given μ, p_0 and σ_2, (σ_1, σ_{1m}) is a best response of player 1 at any state.

2. Given μ, p_0 and σ_1, σ_2 is a best response of player 2 at any state.

3. p_0 is the invariant distribution induced by the Markov chain generated by σ_1 and σ_{1m}, and μ is consistent to the Bayes' rule whenever it is available, i.e.,

$$\mu(s) = \frac{\mu_0 p_1(s)}{\mu_0 p_1(s) + (1-\mu_0) p_0(s)}.[5] \tag{3.1}$$

That is, we focus on stationary equilibria in which equilibrium strategies do not depend on the calendar time, while we allow non-stationary deviations.

We first establish the existence of equilibria.

Proposition 3.1. *There exists a stationary PBE.*

3.3 Characterization of Equilibrium Information Disclosure

From now on we assume $\mu_0 > 0$. We first characterize information disclosure behavior of player 1 on the equilibrium path in any equilibrium. Note that the continuation value is completely determined by disclosed information. This implies that if information is withheld, the continuation value is the same regardless of the actual quality. This immediately implies that high quality product, which involves the sacrifice of static payoff, is made only when the quality will be disclosed.

Proposition 3.2. *For any $\beta \in [0, 1]$,*

$$\sigma_{1m}(H|s, H) = 1$$

for any $s \in \mathcal{S}$ in any equilibrium.

[5] Note that for the commitment type, messages are randomly sent with probability $\beta \in [0, 1]$. Thus, $p_1 \in \Delta(S)$ is exogenously determined.

Proof. Suppose that there is a state $s \in \mathcal{S}$ in which $\sigma_{1m}(B|s, H) > 0$. Consider a deviation to play L then choose B. Since the action is not disclosed, the continuation payoffs should be the same. Then by Assumption 3.1, this is a profitable deviation. Contradiction. □

Note that this result applies even to off-the-equilibrium paths.[6]

3.4 Equilibrium Characterization

3.4.1 Equilibrium Dynamics

Suppose $\beta \in [0, 1)$. One implication of this is that even after observing a bad signal, there is still possibility that he might be the commitment type.[7] Under this assumption, it turns out that the equilibrium dynamics crucially depends on the prior belief. Consider the case with very high prior belief for the commitment type. After observing a succession of bad signals, if the posterior is still high enough to induce the trusting action of player 2, then there would be no point of building reputation. Let

$$\bar{\mu}(\beta, K) \equiv \frac{\bar{\alpha}}{(1-\bar{\alpha})(1-\beta)^K + \bar{\alpha}}. \tag{3.2}$$

Proposition 3.3. *For any $\mu_0 > \overline{\mu}(\beta, K)$, there is a unique equilibrium in which $\sigma_1(L|s) = 1$, $\sigma_{1m}(B|s, a_1) = 1$ and $\sigma_2(h|s) = 1$ for all $s \in \mathcal{S}$ and $a_1 \in A_1$.*

The cutoff given by (3.2) is obtained to satisfy the updated belief after observing a public history only consisting of B is still above $\bar{\alpha}$. Such cutoff is high when the disclosure of high quality is more indicative of the commitment type.

Moreover, if the length of observable public history is longer, i.e., observing bad signals more, the resulting posterior is lower. Hence, we need a higher prior to support this equilibrium. In particular, as β goes to 1 or K goes to infinity, the cutoff converges to 1.

[6] This result still holds when $K = \infty$.

[7] In this sense, this assumption has a similarity to imperfect monitoring or the commitment type with mixed action. However, our model is different with that with imperfect monitoring, since player 1 can choose the signal. In this sense, this model resembles more to the case where the commitment plays mixed action.

Corollary 3.1. *$\bar{\mu}(\beta, K)$ is strictly increasing in both K and β. Especially,*

1. *$\bar{\mu}(1, K) = 1$ and $\bar{\mu}(0, K) = \bar{\alpha}$ for all $K \geq 1$.*

2. *$\bar{\mu}(B, K) \to 1$ as $K \to \infty$.*

Note that in this equilibrium, player 1 attains the highest possible payoff in each period.

Observe also that when $\mu_0 < \overline{\mu}(\beta, K)$, such strategy of player 1 cannot support an equilibrium, simply because player 2 plays l.

The previous proposition suggests that there may be an equilibrium with repeated exploitation of reputation is supported as an equilibrium when the prior belief for the commitment type is sufficiently high. The next result shows that there exists such one. Especially, when the prior belief is *just* below the cutoff, $\bar{\mu}(\beta, K)$, there exists an equilibrium which is after successive reputation, only one building period of reputation is followed on the equilibrium path. See Figure 3.2 to glance the equilibrium dynamics when $K = 2$.

Hereafter, by "exploiting reputation" we mean that player 1 plays L with probability 1 and player 2 chooses h with probability 1 in a period. In addition, by "building reputation" we mean both player 1 and 2 play strictly mixed strategy in a period, ant player 1 discloses the realized action when it is H.

Remark 3.1. Because information disclosure is costless and fully informative on the equilibrium path, when $\beta = 1$, the equilibrium dynamics of this model is essentially the same to Liu (2011) when the cost of information is 0 until K periods and infinite from $K + 1$ (so we omit the analysis of this case). As we allow $\beta < 1$ as well, it gives various equilibrium dynamics, as in Proposition 3.3 and in the following theorem. In particular, the equilibrium dynamics depicted in the theorem is opposite to that of Liu (2011) which is characterized by successive building of reputation for one-time cashing in it.

Given a strategy profile $\sigma \equiv ((\sigma_1, \sigma_{1m}), \sigma_2)$, denote the on-path states by $\mathcal{S}^\sigma \subseteq \mathcal{S}$. Also, given $s \in \mathcal{S}$, let $(s \wedge a_1) \in \mathcal{S}$ be the continuation state.

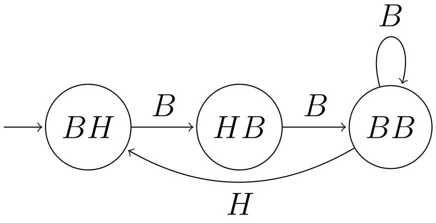


Figure 3.2: The equilibrium (on-path) dynamics described in Proposition 3.4 when $K = 2$

Proposition 3.4. *Let $K \in \mathbb{N}$ and $\beta \in (0,1)$. There exists $\bar{\mu}(\beta, K) \in (0,1)$, $\underline{\mu}(\beta, K) \in (0,1)$ and $\bar{\delta} \in (0,1)$ such that for all $\mu_0 \in [\underline{\mu}(\beta, K), \bar{\mu}(\beta, K)]$ and for all $\delta \geq \bar{\delta}$, there exists an equilibrium $((\sigma_1, \sigma_{1m}), \sigma_2, \mu)$ with K of consecutive exploiting periods followed by a building period on the equilibrium paths, i.e.,*

$$\mathcal{S}^\sigma = \{w_1, w_2, \ldots, w_{K+1}\} \subseteq \mathcal{S}$$

such that

- *K-consecutive exploitation: $\sigma_1(L|w_k) = 1$ and $\sigma_2(h|w_k) = 1$ for any $k = 1, \ldots, K$*
- *1-building: $\sigma_1(H|w_{k+1}) > 0$*
- *Cyclicity: $(w_k \wedge B) = w_{k+1}$ for any $k = 1, \ldots, K$; $(w_{k+1} \wedge H) = w_1$.*

Proof. The proof is constructive. Let us partition $\mathcal{S} = M^K$ based on how remote the most recent H is: let $\mathcal{W}_k$ be the equivalence class when the distance is $k = 1, 2, \ldots, K$, and $\mathcal{W}_{K+1}$ be the class of states which do not have H.[8] Consider the following $((\sigma_1, \sigma_{1m}), \sigma_2, \mu)$:

- Player 1's strategy: for each $s = (s_1, \ldots, s_K) \in \mathcal{S}$

$$\sigma_1(H|s) = \begin{cases} 1 & \text{if } |k : s_k = H| \geq 1 \\ \gamma_1 & \text{o.w.} \end{cases},$$

where $\gamma_1 \in (0,1)$ will be determined shortly (see (3.9)). Also, for any state $s \in \mathcal{S}$, $\sigma_{1m}(H|s, H) = 1$ and $\sigma_{1m}(B|s, L) = 1$. Note that this strategy profile induces on-path

[8] For example, when $K = 4$, $s = BBHB$ this history is classified as $\mathcal{W}_2$.

states of the normal type:

$$\mathcal{S}^\sigma = \{w_1, \ldots, w_K, w_{K+1}\}$$

where $w_1 = (B, \ldots, H), w_2 = (B, \ldots, H, B), \ldots, w_K = (H, B, \ldots, B), w_{K+1} = (H, \ldots, H)$. Note that for each k, $w_k \in \mathcal{W}_k$.

- Player 2's strategy:

$$\sigma_2(h|s) = \begin{cases} 1 & \text{if } s \in \mathcal{W}_k, k = 1, \ldots, K \\ \gamma_2 & \text{if } s \in \mathcal{W}_{K+1} \end{cases}$$

where $\gamma_2 \in (0, 1)$ is to be determined shortly (see (3.15)).[9]

- Belief function and p_0: For any $s = (s_1, \ldots, s_K) \in \mathcal{W}_k$ with $|l : s_l = H| \leq 1$ for some $k = 1, \ldots, K$,

$$\mu(s) = \mu(w_k) := \frac{\mu_0 \beta (1-\beta)^{K-1}}{\mu_0 \beta (1-\beta)^{K-1} + (1-\mu_0) p_0(w_k)} \geq \bar{\alpha} \tag{3.3}$$

and for $s \in \mathcal{W}_0$,

$$\mu(s) = \mu(w_{k+1}) := \frac{\mu_0 (1-\beta)^K}{\mu_0 (1-\beta)^K + (1-\mu_0) p_0(w_{K+1})} \leq \bar{\alpha}. \tag{3.4}$$

and for any $s = (s_1, \ldots, s_K)$ with $|l : s_l = H| > 1$,

$$\mu(s) = 1$$

where $p_0 \in \Delta(\mathcal{S}^\sigma)$ is the invariant distribution induced by (σ_1, σ_{1m}), i.e.,

$$p_0(w_k) = \frac{\gamma_1}{1 + K\gamma_1}, \forall k = 1, \ldots, K \tag{3.5}$$

and

$$p_0(w_0) = \frac{1}{1 + K\gamma_1}. \tag{3.6}$$

Step 1. Consistency of the belief function.

[9] Note that for any state without L and with more than 1 H is off-path of the normal type. For these states, $\sigma_2(s)$=1. In a state with at least 1 L which is off-the-path, σ_2 specifies the same action corresponding to the state which has B in the place of L.

Note that player 2's belief is consistent to the Bayesian update whenever it is available as required in our definition of equilibrium. In particular, for state $s = (s_1, \ldots, s_K)$ with $|k : s_k = H| \geq 1$, if there is no L in s, the long-run player is believed as the commitment type, i.e., $\mu(s) = 1$.

Step 2. Checking incentive of player 2.

First consider off-the-equilibrium states of the normal type, i.e., $s \in \mathcal{S} \setminus \mathcal{S}^\sigma$. Observe that in any state with more than 1 of H, $\mu(s) = 1$; thus, player 2's play of h is clearly incentive compatible.

Consider $s \in \mathcal{W}_{K+1}$. For player 2 to be indifferent,

$$\mu(w_{K+1}) + (1 - \mu(w_{K+1}))\gamma_1 = \bar{\alpha} \tag{3.7}$$

Thus,

$$\mu(w_{K+1}) = \frac{\bar{\alpha} - \gamma_1}{1 - \gamma_1} \tag{3.8}$$

After substituting (3.6) into (3.4), then equating it to (3.8), we obtain,[10]

$$\gamma_1 = \frac{(1 - \mu_0)\bar{\alpha} - \mu_0(1 - \bar{\alpha})(1 - \beta)^K}{(1 - \mu_0) + \mu_0 K(1 - \bar{\alpha})(1 - \beta)^K}. \tag{3.9}$$

Note that $\gamma_1 \geq 0$ if and only if

$$(1 - \mu_0)\bar{\alpha} - \mu_0(1 - \bar{\alpha})(1 - \beta)^K \geq 0$$

Or equivalently,

$$\mu_0 \leq \bar{\mu}(\beta, K) \equiv \frac{\bar{\alpha}}{\bar{\alpha} + (1 - \bar{\alpha})(1 - \beta)^K}. \tag{3.10}$$

Note that $\gamma_1 \in [0, \bar{\alpha}]$ implies $\mu(s_0) \in [0, 1]$ and, therefore, $\mu(w_{K+1}) \leq \bar{\alpha}$ by (3.7). Substitute (3.9) into (3.8), then

$$\mu(w_{K+1}) = \frac{\mu_0(1 - \beta)^K(K\bar{\alpha} + 1)}{\mu_0((K + 1)(1 - \beta)^K - 1) + 1} \tag{3.11}$$

Also by substituting (3.9) into (3.5) and (3.6) we obtain

$$p_0(s) = \frac{(1 - \mu_0)\bar{\alpha} - \mu_0(1 - \bar{\alpha})(1 - \beta)^K}{(1 - \mu_0)(K\bar{\alpha} + 1)}, \forall s \in \{w_1, \ldots, w_K\} \tag{3.12}$$

[10] For the detail of the derivation, see Appendix.

and

$$p_0(w_{K+1}) = \frac{1-\mu_0\left(K\left(\bar{\mu}-1\right)\left(1-\beta\right)^K+1\right)}{\left(1-\mu_0\right)\left(K\bar{\alpha}+1\right)},$$

respectively. Also by substituting (3.12) into (3.3), and requiring it is greater than $\bar{\alpha}$ to rationalize $\sigma_2(s)[h]=1$ in these states, we obtain[11]

$$\mu_0 \geq \underline{\mu}(\beta, K) \equiv \frac{\bar{\alpha}}{\bar{\alpha}+(1-\bar{\alpha})(1-\beta)^K\left(1+(K+\frac{1}{\bar{\alpha}})\frac{\beta}{1-\beta}\right)} \tag{3.13}$$

Clearly $\bar{\mu}(\beta, K) \geq \underline{\mu}(\beta, K)$. Combining (3.10) with (3.13), we obtain the inequality

$$\underline{\mu}(\beta, K) \leq \mu_0 \leq \bar{\mu}(\beta, K).$$

Step 3. Checking incentive of player 1.

Let $V_1 : \mathcal{S} \to \mathbb{R}$ be the continuation value at state s, induced from $((\sigma_1, \sigma_{1m}), \sigma_2, \mu)$. We first check player 1's incentive on the equilibrium paths, i.e., $s \in \{w_1, \ldots, w_{K+1}\}$. From

$$V_1(w_{K+1}) = (1-\delta)(u_1(L,h)\gamma_2 + u_1(L,l)(1-\gamma_2)) + \delta V_1(w_{K+1})$$

we have

$$V_1(w_{K+1}) = \gamma_2 u_1(L,h) + (1-\gamma_2)u_1(L,l). \tag{3.14}$$

It can be easily seen after a few substitutions,

$$\begin{aligned} V_1(w_{K-k}) &= (1-\delta)(1+\delta+\cdots+\delta^k)u_1(L,h) + \delta^{k+1}V_1(w_{K+1}) \\ &= (1-\delta^{k+1})u_1(L,h) + \delta^{k+1}V_1(w_{K+1}) \end{aligned}$$

In particular,

$$V_1(w_1) = (1-\delta^K)u_1(L,h) + \delta^K V_1(w_{K+1})$$

From

$$V_1(w_{K+1}) = (1-\delta)(u_1(H,h)\gamma_2 + (1-\gamma_2)u_1(L,l)) + \delta V_1(w_1)$$

and (3.14), we have[12]

$$\gamma_2 = \frac{(\delta+\cdots+\delta^K)(u_1(L,h)-u_1(L,l)) + (u_1(H,l)-u_1(L,l))}{(u_1(L,h)-u_1(L,l))(1+\delta+\cdots+\delta^K) - (u_1(H,h)-u_1(H,l))} \tag{3.15}$$

[11] For the detail of the derivation, see Appendix.

[12] For the detail of the derivation, see Appendix.

Note that the denominator is strictly positive for all $\delta \in (0,1)$ by Lemma 3.1. In addition, when δ is sufficiently close to 1, the numerator is strictly positive.[13] Moreover, the denominator is always bigger than the numerator.[14] Therefore, γ_2 is well-defined as a probability when δ is sufficiently close to 1.

Since at w_{K+1} player 1 uses a mixed strategy, so we only need to check the other states. The following lemma is a crucial observation. We claim that at w_k, playing L is strictly preferred by player 1. First observe that

$$V_1(w_1) > V_1(w_2) > \cdots > V_1(w_{K+1})$$

because $u_1(L,h) > u_1(L,l)$. Note that for any k, if player 1 plays H then withholding it, then the next period the continuation payoff is $V_1(w_1)$. At w_{K+1} in which player 2 plays h strictly less than 1, playing H and L are indifferent. Since for $k = 1, \ldots, K$, player 2 plays h with probability 1, playing H is strictly preferred to player 1.

Let us consider off-path states. Note that for $s \in \mathcal{W}_k$, $V_1(s) = V_1(w_k)$. To see this, first note that for $s \in \mathcal{W}_{K+1}$,

$$V_1(s) = V_1^{K+1} \equiv (1-\delta)u_1(L, \sigma_2(w_{K+1})) + \delta V_1^{K+1}$$

where we use that at s, $\sigma_2(s) = \sigma_2(w_{K+1})$. Thus,

$$V_1^{K+1} = u_1(L, \sigma_2(w_{K+1})) = V_1(w_{K+1}).$$

Given this, we can easily see that for $s \in \mathcal{W}_K$, $V_1(s) = V_1^K \equiv (1-\delta)u_1(L, \sigma_2(w_K)) + \delta V_1^{K+1}$; but $V_1^{K+1} = V_1(w_{K+1})$, thus $V_1^K = V_1(w_K)$. Similarly we can show that for any $k \in \{1, 2, \ldots, K+1\}$, $V_1^{k+1} = V_1(w_k)$. Then, the incentive compatibility follows since $\sigma_1(s) = \sigma_1(w_k)$ for any $s \in \mathcal{W}_k$ for any $k = 1, \ldots, K+1$. □

[13] By Condition (b) in Assumption 3.1, $u_1(L,h) - u_1(L,l) > u_1(H,h) - u_1(H,l)$. Thus,

$$u_1(L,h) - u_1(L,l) + u_1(H,l) - u_1(L,l) > u_1(H,h) - u_1(H,l) + u_1(H,l) - u_1(L,l) = u_1(H,h) - u_1(L,l).$$

Then, by (e), $u_1(H,h) - u_1(L,l) > 0$.

[14] This is because $u_1(L,h) - u_1(L,l) - (u_1(H,h) - u_1(H,l)) > 0$, while $u_1(H,l) - u_1(L,l) < 0$.

The reason why β strictly less than 1 is necessary for the dynamics in this theorem is obvious: it is to sustain posterior belief higher enough, even after exploiting reputation in a period. Note also that we cannot have this equilibrium when $\mu_0 < \underline{\mu}(\beta, K)$ in that the reputation will be depleted before exploiting K periods. We also obtain a corollary of this theorem, the underlying intuition of which is similar to Corollary 3.1.

Corollary 3.2. *The minimum of $\bar{\mu}(\beta, K)$ and $\underline{\mu}(\beta, K)$ are well-defined and increasing in K.*

Proof. See Appendix. □

This corollary says the prior cannot be too high or too low to have this equilibrium. Also note that $\bar{\mu}(\beta, K)$ of this theorem is exactly the same to the critical value of Proposition 3.3.

Notice that this equilibrium is intuitive and relatively tractable that we can completely characterize for any $K \in \mathbb{N}$. Such tractability comes from the very fact that, except only 1 state, the strategies use pure strategy.

Remark 3.2. Liu (2011) employs a distinct index defined as how many H does a state has until the recent B (in his paper it is L). According to this index, HB and BB should be classified as the same index 0 and therefore player 2 should play the same action. This is not the case generally when $\beta < 1$. Note that $\sigma_2(h|HB) = 1$ and $\sigma_2(h|BB) \in (0, 1)$. First, note that player 1 should have the same continuation payoff regardless of the current play, and therefore only the current incentive affects player 1's action in those states. Suppose $1 \geq \sigma_2(h|HB) > \sigma_2(h|BB) \geq 0$. Unlike $\beta = 1$, it does not imply $\sigma_1(H|HB) > 0$ and $\sigma_1(H|BB) < 1$: even with B in a state, there is still possibility of the commitment type.

Remark 3.3. In the construction in the proof, we use a particular set of off-the-equilibrium beliefs after observing L; namely, we assign the belief at the corresponding on-path state, changing L to B, to an off-path belief at a state s with L. Although our equilibrium notion allows any off-the-equilibrium belief, one might want that the posterior belief after observing L should result in 0 for the commitment type (simply because the commitment type is assumed to never make the low quality; so never disclose L). However, our more permissible

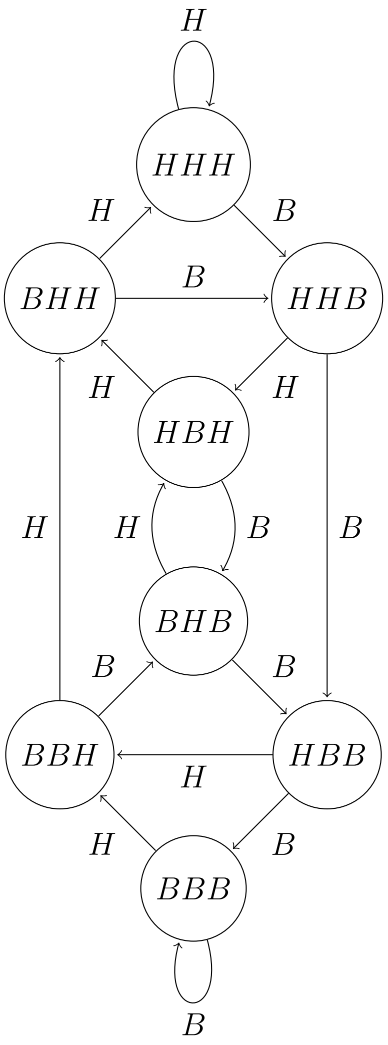


Figure 3.3: Example when $K = 3$

off-path beliefs are also plausible given we consider the possibility of the commitment type making "mistakes."

3.5 Discussion and Conclusion

We conclude this paper by discussing potential future directions which we are considering to pursue.

3.5.1 Costly Verification

In reality, information disclosure is often costly (e.g., having a certification). Suppose that the cost of information disclosure is $c > 0$. We can expect that when c is extremely large, the long-run player would never disclose the quality. It would be interesting to study when c has an intermediate value.

3.5.2 Cheap Talk

In this paper, we study when the message of the long-run player is verifiable. We might also think about situations where the long-run player can send any message, for example, it may send H even the quality of the product is low. Then, a natural question is whether there is an equilibrium in which the cheap talk is informative.

3.5.3 General Prior

At this moment, we have only results when μ is sufficiently large (at least $\bar{\alpha}$). This is imaginable situation; however, probably one may be more interested in when the prior belief for the commitment type is lower (or extremely low "grain of truth").

The problem is that an equilibrium behavior, in principle, may be different for each state; and the number of states is large when K large (see Figure 3.3 for $K = 3$). A natural approach is try to find a smaller set of new "states" which we can use without loss and players' play only depend on these states.

Conjecture 3.1. *When μ_0 is sufficiently low, there is an equilibrium in which the long-run player builds reputation until there is no B or L; then it exploits one time. When building reputation, both player 1 and player 2 are indifferent in their actions.*

At this moment, we do not know exactly how low the prior should be.

3.5.3.1 Player 1's strategy

To specify, player 1's strategy, let us introduce $\mathcal{Y}_k$, $k = 0, \dots, K$.

$$\mathcal{Y}_k := \{s = (s_1, \dots, s_K) \in \mathcal{S} : |\{l : s_l = H\}| = k\}.$$

That is, the set of states which has k of Hs.

Conjecture 3.2. *Suppose for $s, s' \in \mathcal{Y}_k$*

$$\sigma_1(s) = \sigma_1(s') = \gamma_k$$

and $\sigma_{1m}(H|s, a_1) = 1$ if $a_1 = H$; otherwise, $\sigma_{1m}(B|s, a_1) = 1$ for any state s. In addition, suppose that there is a unique invariant distribution $p_0 \in \Delta(\mathcal{S})$. Then,

$$p_0(s) = p_0(s')$$

and

$$\mu(s) = \mu(s').$$

Lemma 3.2. *When $K = 3$, the conjecture is true.*

Proof. It is enough to show that there is an invariant distribution in which $p_0(s) = p_0(s') = p_0^k$ for all $s, s' \in \mathcal{Y}_k$.

Given this restriction,

$$p_0(BBH) = p_0^1 = \gamma_0 p_0(BBBB) + \gamma_1 p(HBB) = \gamma^0 p_0^0 + \gamma^1 p_0^1 \iff p_0^1 = \frac{\gamma_0}{1-\gamma_1} p_0^0$$

$$p_1(BHH) = p_0^2 = \gamma_1 p_0(BBH) + \gamma_1 p_0(HBH) = \gamma^1 p_0^1 + \gamma^2 p_0^2 \iff p_0^2 = \left(\frac{\gamma_1}{1-\gamma_2}\right)\left(\frac{\gamma_0}{1-\gamma_1}\right) p_0^0$$

$$p_0(HHH) = p_0^3 = \gamma_2 p_0(BHH) = \gamma_2 p_0^2 \iff p_0^3 = \gamma_2 \left(\frac{\gamma_1}{1-\gamma_2}\right)\left(\frac{\gamma_0}{1-\gamma_1}\right) p_0^0$$

and

$$p_0^0 + 3p_0^1 + 3p_0^2 + p_0^3 = 1.$$

From this,

$$p_0^0 = \frac{1}{1 + 3\frac{\gamma_0}{1-\gamma_1} + 3\left(\frac{\gamma_1}{1-\gamma_2}\right)\left(\frac{\gamma_0}{1-\gamma_1}\right) + \gamma_2 \left(\frac{\gamma_1}{1-\gamma_2}\right)\left(\frac{\gamma_0}{1-\gamma_1}\right)} \tag{3.16}$$

We need to show that these values satisfy the "preservation equation" for the other states. For example,

$$p_0(BHB) = (1-\gamma_1)p_0(BBH) + (1-\gamma_2)p_0(HBH).$$

From the above

$$\frac{\gamma_0}{1-\gamma_1}p_0 = (1-\gamma_1)\left(\frac{\gamma_0}{1-\gamma_1}\right)p_0^0 + (1-\gamma_2)\left(\frac{\gamma_1}{1-\gamma_2}\right)p_0^0$$

The LHS is

$$\left(\gamma_0 + \frac{\gamma_1\gamma_0}{1-\gamma_1}\right)p_0 = \frac{\gamma_0}{1-\gamma_1}p_0.$$

For HBB, note that the preservation equation is the same. For other states $k \geq 2$, we can use the symmetry. □

Conjecture 3.3. *Consider a strategy profile in which player 1 is indifferent between two actions at any state $s \in \mathcal{Y}_k$ for any $k \geq 1$. Suppose that player 2 is indifferent in these states. Then, there exists an equilibrium in which for states $s, s' \in \mathcal{Y}_k$,*

$$\sigma_1(s) = \sigma_1(s').$$

Lemma 3.3. *For $K = 3$, the previous conjecture is true.*

Proof. Since player 1 is indifferent at s, any play $\sigma_1(H|s) \in [0,1]$ is optimal. Given this we only need to satisfy

$$\mu(s) + (1-\mu(s))\sigma_1(H|s) = \bar{\alpha}$$
$$\mu(s') + (1-\mu(s'))\sigma_1(H|s') = \bar{\alpha}$$

But, we know that $\mu(s) = \mu(s')$, because obviously 1) $p_1(s) = p_1(s')$; and 2) $p_0(s) = p_0(s')$ by Lemma 3.2. In addition, $\sigma_1(H|s) = \sigma_1(H|s')$ implies $p_0(s) = p_0(s')$ (thus also $\mu_0(s) = \mu_0(s')$). □

We do not know whether every equilibrium should satisfy this property. We shall see that this lemma substantially simplifies the problem when $K = 3$ shortly.

Suggested by this conjecture, let us focus on player 1's strategy such that

$$\sigma_1(H|s) = \gamma_k, \forall s \in \mathcal{Y}_k.$$

3.5.3.2 Player 2's Strategy

We focus on equilibria in which player 2's strategy only depends on the location of the most recent B.

Let $\mathcal{W}_k \in (0,1)$ be the set of states whose most recent B or L is k-distant from the current period. Let $\mathcal{W}_{K+1} \equiv \{(H,\ldots,H)\}$. Let

$$\sigma_2(h|s) = \eta_k, \forall s \in \mathcal{W}_k.$$

If there is no B or L, then player 2 plays h with probability 1. Let V_1^k is the continuation value in any state whose most recent B is k-distant from the current period.

$$\begin{aligned} V_1^1 &= (1-\delta)u_1(H,\eta_1) + \delta V_1^2 \\ &= (1-\delta)u_1(L,\eta_1) + \delta V_1^1 \end{aligned}$$

From the second inequality,

$$V_1^1 = u_1(L,\eta_1)$$

Generally, for any $k = 1,\ldots,K-1$,

$$\begin{aligned} V_1^k &= (1-\delta)u_1(H,\eta_k) + \delta V_1^{k+1} \\ &= (1-\delta)u_1(L,\eta_k) + \delta V_1^1 \end{aligned}$$

and

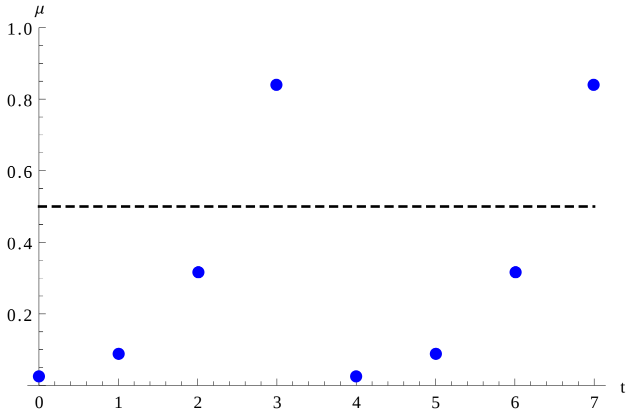
$$V_1^{K+1} = (1-\delta)u_1(L,h) + \delta V_1^1.$$

Conjecture 3.4. *For sufficiently large δ, there exists $\eta_k \in [0,1]$, $k = 1,\ldots,K$ which satisfy the equations. In addition,*

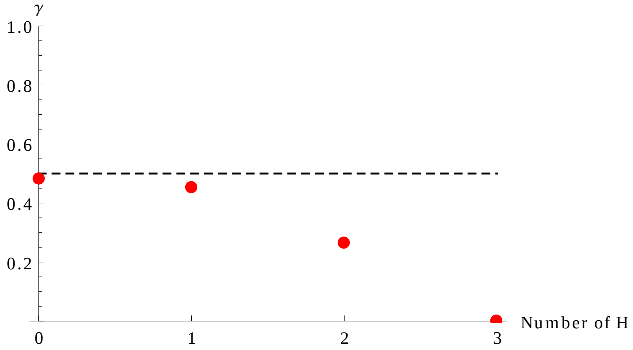
$$V_1^{K+1} > V_1^K > \cdots > V_1^1.$$

3.5.3.3 Example: $K = 3$

Although we have not been able to establish results for generic K, we know the conjectures are true for $K = 3$.



(a) Reputation dynamics. At time 0, state is BBB



(b) Behavior of player 1 (probability of playing H) depending on the number of H in s

Figure 3.4: When $\bar{\alpha} = 1/2$, $\beta = 3/4$, $\mu_0 = 1/4$ and $K = 3$

Proposition 3.5. *For $K = 3$, Conjecture 3.1 is true for some μ_0.*

See Figure 3.4 for the reputation dynamics and equilibrium of the long-run player in this equilibrium.

Let us first explicitly obtain player 1's strategy. As we discussed, we are focusing on equilibria in which for any states in the same $\mathcal{Y}_k$, player 1 plays the same. Then, by Lemma 3.2, we know that $p_0(s) = p_0(s')$ for all $s, s' \in \mathcal{Y}_k$ for any k.

In addition, we are focusing on equilibria where player 2 is indifferent at any $s \in \mathcal{Y}_k$ for some $k = 1, \ldots, K-1$; and $\sigma_2(h|s) = 1$ for $s = HHH$. Thus, given $k \in \{1, \ldots, K-1\}$,

$$\mu^k + (1-\mu^k)\gamma_k = \bar{\alpha}, \forall s \in \mathcal{Y}_k$$

where

$$\begin{aligned}\mu^k &= \frac{\mu_0 p_1^k}{\mu_0 p_1^k + (1-\mu_0)p_0^k} \\ p_1^k &= \beta^k(1-\beta)^{K-k}\end{aligned}$$

and p_0^k are determined by (3.16). Thus, we have 3 equations for 3 unknowns (γ_0, γ_1 and γ_2). To be more explicit,

$$\mu_0 p_1^k + (1-\mu_0)p_0^k\gamma_k = \bar{\alpha}(\mu_0 p_1^k + (1-\mu_0)p_0^k)$$

$$\iff p_0^k(\gamma_k - \bar{\alpha}) = \frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)p_1^k$$

That is,

$$\begin{aligned}\frac{1}{D_\gamma}(\gamma_0 - \bar{\alpha}) &= \frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)\beta^0(1-\beta)^3 \\ \frac{\frac{\gamma_0}{1-\gamma_1}}{D_\gamma}(\gamma_1 - \bar{\alpha}) &= \frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)\beta(1-\beta)^2 \\ \frac{\left(\frac{\gamma_1}{1-\gamma_2}\right)\left(\frac{\gamma_0}{1-\gamma_1}\right)}{D_\gamma}(\gamma_2 - \bar{\alpha}) &= \frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)\beta^2(1-\beta)\end{aligned}$$

where D_γ is the denominator of (3.16).

Proposition 3.6. *Suppose $\beta > 1/2$. Then we have*

$$\gamma_0 > \gamma_1 > \gamma_2.$$

In addition we have explicit expression for γ_0:

$$\gamma_0 = \bar{\alpha} + \frac{\frac{\mu_0}{1-\mu_0}(1-\beta)^3 \left(\bar{\alpha} - 1 - \bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)}{1 + \frac{\mu_0}{1-\mu_0}(1-\beta)^3 \frac{1}{1-\bar{\alpha}}\Phi}$$

where

$$\Phi \equiv 3 + \bar{\alpha}^2 + 3(1-\bar{\alpha})\frac{\beta}{1-\beta} + (3 + \bar{\alpha} + \beta - 5\bar{\alpha}\beta)\frac{\beta}{(1-\beta)^2}.$$

Proof. See Appendix. □

To obtain player 2's strategy at states in $\mathcal{W}_k$ for $k = 1, 2, 3$ are determined so that player 1 is indifferent:

$$\begin{aligned} V_1^2 &= (1-\delta)u_1(H, \eta^2) + \delta V_1^3 \\ &= (1-\delta)u_1(L, \eta^2) + \delta V_1^1 \end{aligned}$$

$$\begin{aligned} V_1^3 &= (1-\delta)u_1(H, \eta^3) + \delta V_1(HHH) \\ &= (1-\delta)u_1(L, \eta^3) + \delta V_1^1 \end{aligned}$$

From

$$V_1(HHH) = (1-\delta)u_1(L, h) + \delta V_1^1$$

Thus,

$$V_1(HHH) = (1-\delta)u_1(L, h) + \delta u_1(L, \eta^1)$$

Proposition 3.7. *When $\delta \in [0, 1)$ sufficiently high,*

$$\eta_1 < \eta_2 < \eta_3$$

and we have explicit expression for η_k for $k = 1, 2, 3$:

$$\begin{aligned} \eta_3 &= \frac{\xi + \xi^2 + \xi^3 - \frac{\underline{d}}{\bar{d}-\underline{d}}}{1 + \xi + \xi^2 + \xi^3} \\ \eta_2 &= \eta_3 - (1-\eta_3)\xi \\ \eta_1 &= \eta_3 - (1-\eta_3)\xi - (1-\eta_3)\xi^2. \end{aligned}$$

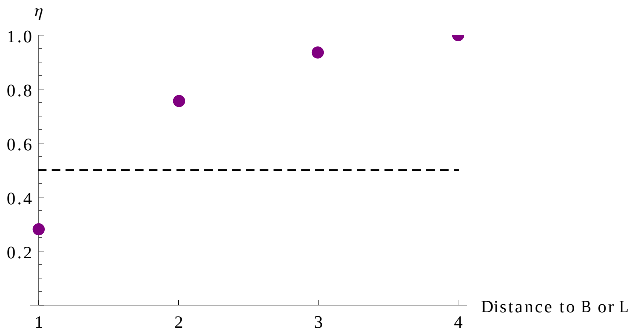


Figure 3.5: Player 2's behavior (probability of playing h)

where

$$\bar{d} \equiv u_1(L,h) - u_1(H,h)$$
$$\underline{d} \equiv u_1(L,l) - u_1(H,l)$$
$$z \equiv u_1(L,h) - u_1(L,l)$$

and

$$\xi \equiv \delta \frac{z}{\bar{d} - \underline{d}}$$

See Figure 3.5.

Proof. See Appendix. □

3.6 Appendix

3.6.1 Omitted Proofs

3.6.1.1 Proof of Proposition 3.1

Proof. Fix $p_1 \in \Delta(\mathcal{S})$. Define a point-to-set correspondence[15]

$$(BR_1, BR_{1m}, BR_2, m, q_0) : (\Delta(A_1) \times \Delta(M) \times \Delta(A_2) \times [0,1])^{|S|} \times \Delta(\mathcal{S})$$

[15] See Rosenthal (1979) for a proof of existence of equilibrium in repeated games with random matching. Our proof is hinted by this. Also, for Kakutani fixed theorem, we refer to an exposition in Myerson (1991).

$$\rightrightarrows (\Delta(A_1) \times \Delta(A_2) \times [0,1])^{|\mathcal{S}|} \times \Delta(\mathcal{S})$$

where, for $\sigma_1 \in (\Delta(A_1))^{|\mathcal{S}|}$, $\sigma_{1m} \in (\Delta(M))^{|\mathcal{S}|}$, $\sigma_2 \in (\Delta(A_2))^{|\mathcal{S}|}, p_0 \in \Delta(\mathcal{S}), \mu \in [0,1]^{|\mathcal{S}|}$, $BR_1^s(\sigma_1, \sigma_{1m}, \sigma_2, \mu, p_0)$, $BR_{1m}^s(\sigma_1, \sigma_{1m}, \sigma_2, \mu, p_0)$ and $BR_2^s(\sigma_1, \mu, p_0)$ are the best response correspondence at state $s \in \mathcal{S}$; i.e.,

$$BR_1^s(\sigma_1, \sigma_{1m}, \sigma_2, p_0) := \operatorname*{arg\,max}_{\alpha_1 \in \Delta(A_1)} (1-\delta) u_1(\alpha_1, \sigma_2(s)) + \delta \sum_{a_1 \in A_1} \sum_{m \in M} V(s \wedge m) \sigma_{1m}(m|s, a_1) \alpha_1(a_1)$$

$$BR_{1m}^s(\sigma_1, \sigma_{1m}, \sigma_2) := \operatorname*{arg\,max}_{\alpha_{1m} \in \Delta(M)} \sum_{a_1 \in A_1} \sum_{m \in M} V(s \wedge m) \alpha_{1m}(m|s, a_1) \sigma_1(a_1|s)$$

where $V : \mathcal{S} \to \mathbb{R}$ such that for each s

$$V(s) = (1-\delta) u_1(\sigma_1(s), \sigma_2(s)) + \delta \sum_{a_1 \in A_1} \sum_{m \in M} V(s \wedge m) \sigma_{1m}(m|s, a_1) \sigma_1(a_1|s).$$

Also

$$BR_2^s(\sigma_1, \mu, p_0) = \operatorname*{arg\,max}_{\alpha_2' \in \Delta(A_2)} \mu(s) u_1(H, \alpha_2') + (1-\mu(s)) u_2(\sigma_1(s), \alpha_2'),$$

$$q_0^s(p_0, \sigma_1, \sigma_{1m}) = \sum_{s' \in S, m \in M : (s' \wedge m) = s} p_0^{s'} \sum_{a_1 \in A_1} \sigma_{1m}(m|s', a_1) \sigma_1(a_1|s')$$

and lastly,

$$m^s(p_0) = \frac{\mu p_1(s)}{\mu p_1(s) + (1-\mu) p_0(s)}$$

if either $p_1(s) > 0$ or $p_0(s) > 0$; otherwise $m^s(p_0) = [0,1]$.

Clearly the domain is nonempty, convex and compact. The best response at each $s \in S$ is convex-valued, since it is a subset of $\Delta(A_i)$ and defined by a linear inequality. Also by the continuity of expected utility, it is also upper-hemicontinuous. For the similar reason q_0^s is convex-valued, and clearly it is upper-hemicontinuous, and for the similar reason, m^s is so. Since a product of upper-hemicontinuous correspondence is upper-hemicontinuous, by Kakutani fixed-point theorem, there is an equilibrium. □

3.6.1.2 Proof of Proposition 3.3

Proof. In the conjectured equilibrium,

$$\frac{\mu_0 (1-\beta)^K}{\mu_0 (1-\beta)^K + (1-\mu_0) 1} \geq \bar{\alpha}$$

$$\Leftrightarrow \mu_0 \geq \frac{\bar{\alpha}}{(1-\bar{\alpha})(1-\beta)^K + \bar{\alpha}}.$$

This implies that given $\beta \in [0,1)$ if $\mu_0 \geq \overline{\mu}(\beta, K) \equiv \frac{\bar{\alpha}}{(1-\bar{\alpha})(1-\beta)^K+\bar{\alpha}}$ then player 2 chooses h even though he believes that the normal type always chooses L. Thus, the deviation to H is not profitable.

Next, we argue that this is a unique equilibrium. Suppose that there is an equilibrium in which player 1 plays H in a history. Consider the deviation to L then withhold this information in any following history. Since $\mu_0 > \bar{\mu}(\beta, K)$, player 2 plays h in these history and it gives $u_1(L,h)$ which is the maximum payoff player 1 could obtain. So it is a profitable deviation. Contradiction. □

3.6.1.3 Proof of Corollary 3.2

Proof. From equation (3.13), in order to show $\underline{\mu}(\beta, K)$ is increasing in K, it is sufficient to $(1-\beta)^K \left(1 + (K + \frac{1}{\bar{\alpha}})\frac{\beta}{1-\beta}\right) = (1-\beta)^K + (K+\frac{1}{\bar{\alpha}})\beta(1-\beta)^{K-1}$ is decreasing. The derivative w.r.t. K is

$$\begin{aligned}\log(1-\beta)(1-\beta)^K + \beta(1-\beta)^{K-1} + \left(K + \frac{1}{\bar{\alpha}}\right)\beta(1-\beta)^{K-1}\log(1-\beta) \\ = (1-\beta)^{K-1}\left[\log(1-\beta)(1-\beta) + \beta + \left(K+\frac{1}{\bar{\alpha}}\right)\beta\log(1-\beta)\right]\end{aligned}$$

Note that $\log(1-\beta) < 0$, thus

$$\log(1-\beta)\left[(1-\beta) + \left(K + \frac{1}{\bar{\alpha}}\right)\beta\right] + \beta < \log(1-\beta) + \beta < 0$$

since $\beta < 1$ and $\log(1-\beta)+\beta$ is strictly decreasing when $\beta \in (0,1)$ and $\log 1 - 1 = 0$. For the part for $\bar{\mu}(\beta, K)$ note that the value is the same to that of Proposition 3.3. □

3.6.1.4 Proof of Proposition 3.6

Proof. The above conditions are equivalent to

$$\gamma_k = \bar{\alpha} - D_\gamma\left(\frac{\mu_0}{1-\mu_0}\right)(1-\bar{\alpha})\beta^k(1-\beta)^{K-k}, \forall k \in \{0,1,2\}.$$

Note that

$$\beta^k(1-\beta)^{K-k} = (1-\beta)^K \left(\frac{\beta}{1-\beta}\right)^k.$$

Thus, as long as $\beta > 1/2$, γ_k is increasing in k.

In order to obtain the explicit expression in the statement, note that

$$\begin{aligned}
& \frac{\gamma_0-\bar{\alpha}}{\gamma_1-\bar{\alpha}}\frac{1-\gamma_1}{\gamma_0} = \frac{1-\beta}{\beta} \\
\iff & \frac{\gamma_1-\bar{\alpha}}{1-\gamma_1} = \frac{\beta}{1-\beta}\frac{\gamma_0-\bar{\alpha}}{\gamma_0} \\
\iff & \gamma_1-\bar{\alpha} = (1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}(1-\gamma_1) \\
\iff & \gamma_1\left(1+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}\right) = \bar{\alpha}+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta} \\
\iff & \gamma_1 = \frac{\bar{\alpha}+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}}{1+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}} = 1+\frac{\bar{\alpha}-1}{1+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}}
\end{aligned}$$

For future reference,

$$1-\gamma_1 = \frac{1-\bar{\alpha}}{1+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}}$$

Thus,

$$\frac{\gamma_1}{1-\gamma_1} = \frac{\bar{\alpha}+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}}{1-\bar{\alpha}}$$

Also,

$$\frac{1}{1-\gamma_1} = \frac{1+\left(1-\frac{\bar{\alpha}}{\gamma_0}\right)\frac{\beta}{1-\beta}}{1-\bar{\alpha}}$$

Similarly,

$$\frac{\gamma_2}{1-\gamma_2} = \frac{\bar{\alpha}+\left(1-\frac{\bar{\alpha}}{\gamma_1}\right)\frac{\beta}{1-\beta}}{1-\bar{\alpha}}$$

and

$$\frac{1}{1-\gamma_2} = \frac{1+\left(1-\frac{\bar{\alpha}}{\gamma_1}\right)\frac{\beta}{1-\beta}}{1-\bar{\alpha}}$$

Note that

$$G_\gamma = 1+3\gamma_0\frac{1}{1-\gamma_1}+3\gamma_0\frac{1}{1-\gamma_2}\frac{\gamma_1}{1-\gamma_1}+\gamma_0\frac{\gamma_2}{1-\gamma_2}\frac{\gamma_1}{1-\gamma_1}$$

Note that

$$\begin{aligned}
\frac{\gamma_1}{1-\gamma_2} &= \frac{\gamma_1\left(1+\left(1-\frac{\bar{\alpha}}{\gamma_1}\right)\frac{\beta}{1-\beta}\right)}{1-\bar{\alpha}} \\
&= \frac{\gamma_1+\left(\gamma_1\frac{\beta}{1-\beta}-\bar{\alpha}\frac{\beta}{1-\beta}\right)}{1-\bar{\alpha}} \\
&= \frac{\gamma_1(1+\frac{\beta}{1-\beta})-\bar{\alpha}\frac{\beta}{1-\beta}}{1-\bar{\alpha}}
\end{aligned}$$

From this, the third term is

$$\begin{aligned}
\frac{\gamma_1}{1-\gamma_2}\frac{1}{1-\gamma_1} &= \left(\frac{1+\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right)\left(\frac{\gamma_1}{1-\gamma_1}\right)-\frac{\bar{\alpha}}{1-\bar{\alpha}}\frac{\beta}{1-\beta}\left(\frac{1}{1-\gamma_1}\right) \\
&= \left(\frac{1+\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right)\left(\frac{\bar{\alpha}+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right)-\frac{\bar{\alpha}}{1-\bar{\alpha}}\frac{\beta}{1-\beta}\left(\frac{1+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right) \\
&= \frac{\bar{\alpha}+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}-\bar{\alpha}\beta-\bar{\alpha}\beta(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)} \\
&= \frac{\bar{\alpha}(1-\beta)+(1-\frac{\bar{\alpha}}{\gamma_0})(1-\bar{\alpha}\beta)\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)}
\end{aligned}$$

To obtain the last term,

$$\begin{aligned}
\frac{\gamma_2}{1-\gamma_2}\gamma_1 &= \left(\frac{\bar{\alpha}+\left(1-\frac{\bar{\alpha}}{\gamma_1}\right)\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right)\gamma_1 \\
&= \frac{\bar{\alpha}\gamma_1+(\gamma_1-\alpha)\frac{\beta}{1-\beta}}{1-\bar{\alpha}} \\
&= \frac{(\bar{\alpha}+\frac{\beta}{1-\beta})\gamma_1-\bar{\alpha}\frac{\beta}{1-\beta}}{1-\bar{\alpha}}
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{\gamma_2}{1-\gamma_2}\frac{\gamma_1}{1-\gamma_1} &= \frac{\bar{\alpha}+\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\left(\frac{\gamma_1}{1-\gamma_1}\right)-\frac{\bar{\alpha}}{1-\bar{\alpha}}\frac{\beta}{1-\beta}\left(\frac{1}{1-\gamma_1}\right) \\
&= \frac{\bar{\alpha}+\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\left(\frac{\bar{\alpha}+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right)-\frac{\bar{\alpha}}{1-\bar{\alpha}}\frac{\beta}{1-\beta}\left(\frac{1+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right) \\
&= \frac{\bar{\alpha}(1-\beta)+\beta)(\bar{\alpha}+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta})}{(1-\bar{\alpha})^2(1-\beta)}-\frac{\bar{\alpha}\beta(1+(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta})}{(1-\bar{\alpha})^2(1-\beta)} \\
&= \frac{\bar{\alpha}(\bar{\alpha}(1-\beta)+\beta)-\bar{\alpha}\beta+(\bar{\alpha}(1-\beta)+\beta-\bar{\alpha}\beta)(1-\frac{\bar{\alpha}}{\gamma_0})\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)}
\end{aligned}$$

From all of theses,

$$\begin{aligned}
D_\gamma &= 1 + 3\frac{\gamma_0}{1-\gamma_1} + 3\frac{\gamma_1}{1-\gamma_2}\frac{\gamma_0}{1-\gamma_1} + \gamma_2\frac{\gamma_1}{1-\gamma_2}\frac{\gamma_0}{1-\gamma_1} \\
&= 1 + 3\left(\frac{\gamma_0 + (\gamma_0-\bar{\alpha})\frac{\beta}{1-\beta}}{1-\bar{\alpha}}\right) + 3\left(\frac{\bar{\alpha}(1-\beta)\gamma_0 + (\gamma_0-\bar{\alpha})(1-\bar{\alpha}\beta)\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)}\right) \\
&\quad + \frac{\bar{\alpha}^2(1-\beta)\gamma_0 + (\bar{\alpha}(1-\beta)+\beta-\bar{\alpha}\beta)(\gamma_0-\bar{\alpha})\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)} \\
&= 1 + \frac{3\gamma_0(1-\bar{\alpha})(1-\beta) + 3(\gamma_0-\bar{\alpha})(1-\bar{\alpha})\beta + 3\bar{\alpha}(1-\beta)\gamma_0 + 3(\gamma_0-\bar{\alpha})(1-\bar{\alpha}\beta)\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)} \\
&\quad + \frac{\bar{\alpha}^2(1-\beta)\gamma_0 + (\bar{\alpha}(1-\beta)+\beta-\bar{\alpha}\beta)(\gamma_0-\bar{\alpha})\frac{\beta}{1-\beta}}{(1-\bar{\alpha})^2(1-\beta)} \\
&= 1 + \frac{\gamma_0(3+\bar{\alpha}^2)}{(1-\bar{\alpha})^2} + \frac{(\gamma_0-\bar{\alpha})(3(1-\bar{\alpha})\beta + 3(1-\bar{\alpha}\beta)\frac{\beta}{1-\beta} + (\bar{\alpha}(1-\beta)+\beta-\bar{\alpha}\beta)\frac{\beta}{1-\beta})}{(1-\bar{\alpha})^2(1-\beta)} \\
&= 1 + (\gamma_0-\bar{\alpha})\left[\frac{3+\bar{\alpha}^2}{(1-\bar{\alpha})^2} + \frac{(3(1-\bar{\alpha})\beta + 3(1-\bar{\alpha}\beta)\frac{\beta}{1-\beta} + (\bar{\alpha}(1-\beta)+\beta-\bar{\alpha}\beta)\frac{\beta}{1-\beta})}{(1-\bar{\alpha})^2(1-\beta)}\right] \\
&\quad + \bar{\alpha}\frac{3+\bar{\alpha}^2}{(1-\bar{\alpha})^2} \\
&= 1 + (\gamma_0-\bar{\alpha})\frac{1}{(1-\bar{\alpha})^2}\left[3+\bar{\alpha}^2 + 3(1-\bar{\alpha})\frac{\beta}{1-\beta} + (3+\bar{\alpha}+\beta-5\bar{\alpha}\beta)\frac{\beta}{(1-\beta)^2}\right] + \bar{\alpha}\frac{3+\bar{\alpha}^2}{(1-\bar{\alpha})^2}
\end{aligned}$$

Our original equation is

$$\gamma_0 - \bar{\alpha} = \frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)(1-\beta)^3 D_\gamma \tag{3.17}$$

$$\iff \gamma_0 - \bar{\alpha} = \frac{\mu_0}{1-\mu_0}(1-\beta)^3\left((\bar{\alpha}-1) - (\gamma_0-\bar{\alpha})\frac{1}{1-\bar{\alpha}}\Phi - \bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)$$

$$\iff (\gamma_0-\bar{\alpha})\left(1 + \frac{\mu_0}{1-\mu_0}(1-\beta)^3\frac{1}{1-\bar{\alpha}}\Phi\right) = \frac{\mu_0}{1-\mu_0}(1-\beta)^3\left(\bar{\alpha} - 1 - \bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)$$

where

$$\Phi \equiv 3 + \bar{\alpha}^2 + 3(1-\bar{\alpha})\frac{\beta}{1-\beta} + (3+\bar{\alpha}+\beta-5\bar{\alpha}\beta)\frac{\beta}{(1-\beta)^2}$$

Note that $\Phi > 0$, by observing $3+\bar{\alpha}+\beta-5\bar{\alpha}\beta$ because $3 \geq 3\bar{\alpha}\beta$ and $\bar{\alpha}, \beta \geq \bar{\alpha}\beta$.

$$\gamma_0 = \bar{\alpha} + \frac{\frac{\mu_0}{1-\mu_0}(1-\beta)^3\left(\bar{\alpha} - 1 - \bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)}{1 + \frac{\mu_0}{1-\mu_0}(1-\beta)^3\frac{1}{1-\bar{\alpha}}\Phi} \tag{3.18}$$

Note that $\gamma_0 \leq \bar{\alpha}$ as we expected.

From (3.17) and (3.18), we can derive

$$D_\gamma = \frac{\frac{\frac{\mu_0}{1-\mu_0}(1-\beta)^3\left(\bar{\alpha}-1-\bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)}{1+\frac{\mu_0}{1-\mu_0}(1-\beta)^3\frac{1}{1-\bar{\alpha}}\Phi}}{\frac{\mu_0}{1-\mu_0}(\bar{\alpha}-1)(1-\beta)^3} = \frac{\frac{\left(\bar{\alpha}-1-\bar{\alpha}\frac{3+\bar{\alpha}^2}{1-\bar{\alpha}}\right)}{1+\frac{\mu_0}{1-\mu_0}(1-\beta)^3\frac{1}{1-\bar{\alpha}}\Phi}}{(\bar{\alpha}-1)}.$$

□

3.6.1.5 Proof of Proposition 3.7

Proof. Given $\eta_1, \ldots, \eta_K$, we can obtain V_1^k.

$$\begin{aligned} V_1^3 &= (1-\delta)u_1(H,\eta^3) + \delta((1-\delta)u_1(L,h) + \delta u_1(L\eta^1)) \\ &= (1-\delta)u_1(L,\eta^3) + \delta u_1(L,\eta^1) \end{aligned}$$

From this,

$$u_1(L,\eta^3) - u_1(H,\eta^3) = \delta(u_1(L,h) - u_1(L,\eta^1))$$

Similarly,

$$u_1(L,\eta^2) - u_1(H,\eta^2) = \delta(u_1(L,\eta^3) - u_1(L,\eta^1))$$

$$u_1(L,\eta^1) - u_1(H,\eta^1) = \delta(u_1(L,\eta^2) - u_1(L,\eta^1))$$

Also,

$$\eta_1 u_1(L,h) + (1-\eta)u_1(L,l) - (\eta_1 u_1(H,h) + (1-\eta_1)u_1(H,l))$$

$$= \eta_1(u_1(L,h) - u_1(H,h)) + (1-\eta_1)(u_1(L,l) - u_1(H,l))$$

Also,

$$\begin{aligned} u_1(L,\eta_2) - u_1(L,\eta_1) &= \eta_2 u_1(L,h) + (1-\eta_2)u_1(L,l) - (\eta_1 u_1(L,h) + (1-\eta_1)u_1(L,l)) \\ &= (\eta_2-\eta_1)u_1(L,h) + (1-\eta_2-1+\eta_1)u_1(L,l) \\ &= (\eta_2-\eta_1)u_1(L,h) - (\eta_2-\eta_1)u_1(L,l) \\ &= (\eta_2-\eta_1)(u_1(L,h) - u_1(L,l)) \end{aligned}$$

Thus,

$$\eta_1(u_1(L,h) - u_1(H,h)) + (1-\eta_1)(u_1(L,l) - u_1(H,l)) = \delta(\eta_2-\eta_1)(u_1(L,h) - u_1(L,l))$$

Let

$$\bar{d} \equiv u_1(L,h) - u_1(H,h)$$
$$\underline{d} \equiv u_1(L,l) - u_1(H,l)$$
$$z \equiv u_1(L,h) - u_1(L,l)$$

Using these notations,

$$\eta_1\bar{d} + (1-\eta_1)\underline{d} = \delta(\eta_2 - \eta_1)z$$

$$\eta_2 x + (1-\eta_2)y = \delta(\eta_3 - \eta_1)z$$

$$\eta_3\bar{d} + (1-\eta_3)\underline{d} = \delta(1 - \eta_1)z$$

$$\iff \eta_3 = \frac{\delta(1-\eta_1)z - \underline{d}}{\bar{d} - \underline{d}} \tag{3.19}$$

Similarly,

$$\eta_2 = \frac{\delta(\eta_3 - \eta_1)z - \underline{d}}{\bar{d} - \underline{d}}$$
$$\eta_1 = \frac{\delta(\eta_2 - \eta_1)z - \underline{d}}{\bar{d} - \underline{d}}$$

From these,

$$\eta_3 - \eta_2 = \frac{\delta(1-\eta_3)z}{\bar{d} - \underline{d}}$$
$$\eta_2 - \eta_1 = \frac{\delta(\eta_3 - \eta_2)z}{\bar{d} - \underline{d}}$$

Let

$$\xi \equiv \delta\frac{z}{\bar{d} - \underline{d}}$$

Then,

$$\eta_3 - \eta_2 = (1-\eta_3)\eta$$

$$\iff \eta_2 = \eta_3 - (1-\eta_3)\xi \tag{3.20}$$

Also,

$$\eta_2 - \eta_1 = (\eta_3 - \eta_2)\xi.$$

Substituting (3.20),

$$(\eta_3 - (1 - \eta_3)\xi - \eta_1)$$

$$= (\eta_3 - (\eta_3 - (1 - \eta_3)\xi))\xi = (1 - \eta_3)\xi^2.$$

Thus,

$$\eta_1 = \eta_3 - (1 - \eta_3)\xi - (1 - \eta_3)\xi^2. \tag{3.21}$$

By substituting (3.20) and (3.21) into (3.19),

$$\begin{aligned}\eta_3 &= \frac{\delta(1 - (\eta_3 - (1 - \eta_3)\xi - (1 - \eta_3)\xi^2))z - \underline{d}}{\bar{d} - \underline{d}} \\ &= (1 - (\eta_3 - (1 - \eta_3)\xi - (1 - \eta_3)\xi^2)\xi - \frac{\underline{d}}{\bar{d} - \underline{d}}\end{aligned}$$

$$\begin{aligned}\iff \eta_3 &= \xi - \eta_3\xi + (1 - \eta_3)\xi^2 + (1 - \eta_3)\xi^3 - \frac{\underline{d}}{\bar{d} - \underline{d}} \\ &= \xi - \eta^3\xi - \eta_3\xi^2 - \eta_3\xi^3 + \xi^2 + \xi^3 - \frac{\underline{d}}{\bar{d} - \underline{d}}\end{aligned}$$

Or

$$\eta_3(1 + \xi + \xi^2 + \xi^3) = \xi + \xi^2 + \xi^3 - \frac{\underline{d}}{\bar{d} - \underline{d}}$$

$$\iff \eta_3 = \frac{\xi + \xi^2 + \xi^3 - \frac{\underline{d}}{\bar{d} - \underline{d}}}{1 + \xi + \xi^2 + \xi^3}.$$

□

3.6.2 Detailed Derivation of Equations in Proof of Proposition 3.4

3.6.2.1 Derivation of (3.9)

In detail,

$$
\begin{aligned}
&\frac{\mu_0(1-\beta)^K(1+K\gamma_1)}{\mu_0(1-\beta)^K(1+K\gamma_1)+(1-\mu_0)} = \frac{\bar{\alpha}-\gamma_1}{1-\gamma_1} \\
\iff\quad & \mu_0(1-\beta)^K(1+K\gamma_1)(1-\gamma_1) = (\mu_0(1-\beta)^K(1+K\gamma_1)+(1-\mu_0))(\bar{\alpha}-\gamma_1) \\
\iff\quad & \mu_0(1-\beta)^K(1+K\gamma_1)(1-\bar{\alpha}) = (1-\mu_0)(\bar{\alpha}-\gamma_1) \\
\iff\quad & \gamma_1((1-\bar{\alpha})K\mu_0(1-\beta)^K+(1-\mu_0)) = \bar{\alpha}(1-\mu_0)-(1-\bar{\alpha})\mu_0(1-\beta)^K.
\end{aligned}
$$

3.6.2.2 Derivation of (3.13)

$$
\begin{aligned}
\mu(s) = \frac{\mu_0 p_1(s)}{\mu_0 p_1(s)+(1-\mu_0)p_0(s)} &= \frac{1}{1+\frac{1-\mu_0}{\mu_0}\frac{p_0(s)}{p_1(s)}} \geq \bar{\alpha} \\
\iff\quad \frac{1-\bar{\alpha}}{\bar{\alpha}}\frac{\mu_0}{1-\mu_0} &\geq \frac{p_0(s)}{p_1(s)} \\
\iff\quad \frac{1-\bar{\alpha}}{\bar{\alpha}}\frac{\mu_0}{1-\mu_0} &\geq \frac{1}{(1-\mu_0)(K\bar{\alpha}+1)}\frac{(1-\mu_0)\bar{\alpha}-\mu_0(1-\alpha)(1-\beta)^K}{\beta(1-\beta)^{K-1}} \\
\iff\quad \frac{1-\bar{\alpha}}{\bar{\alpha}}\frac{\mu_0}{1-\mu_0}\frac{(1+K\bar{\alpha})}{\bar{\alpha}}\beta(1-\beta)^{K-1} &\geq 1-\frac{\mu_0}{1-\mu_0}\frac{1-\bar{\alpha}}{\bar{\alpha}}(1-\beta)^K \\
\iff\quad \frac{\mu_0}{1-\mu_0}\frac{1-\bar{\alpha}}{\bar{\alpha}}(1-\beta)^K\left(1+\left(K+\frac{1}{\bar{\alpha}}\right)\frac{\beta}{1-\beta}\right) &\geq 1 \qquad (3.22)
\end{aligned}
$$

3.6.2.3 Derivation of (3.15)

$$
\begin{aligned}
&\gamma_2((u_1(L,h)-u_1(L,l))(1-\delta^{K+1})-(1-\delta)(u_1(H,h)-u_1(H,l))) \\
&= (1-\delta)u_1(H,l)+\delta((1-\delta^K)u_1(L,h))+(\delta^{K+1}-1)u_1(L,l) \\
&\Rightarrow \gamma_2 = \frac{(1-\delta)u_1(H,l)+\delta((1-\delta^K)u_1(L,h))+(\delta^{K+1}-1)u_1(L,l)}{(u_1(L,h)-u_1(L,l))(1-\delta^{K+1})-(1-\delta)(u_1(H,h)-u_1(H,l))} \\
&= \frac{u_1(H,l)+\delta(1+\delta+\cdots+\delta^{K-1})u_1(L,h)-(1+\delta+\cdots+\delta^K)u_1(L,l)}{(u_1(L,h)-u_1(L,l))(1+\delta+\cdots+\delta^K)-(u_1(H,h)-u_1(L,l))}
\end{aligned}
$$

Bibliography

Artemov, G., T. Kunimoto, and R. Serrano (2013). Robust virtual implementation: Toward a reinterpretation of the wilson doctrine. *Journal of Economic Theory 148*(2), 424–447.

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics 4*(6), 1236–1239.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society 55*(1), 1–18.

Aumann, R. J. (1998). Common priors: A reply to gul. *Econometrica 66*(4), 929–938.

Bergemann, D. and S. Morris (2005). Robust mechanism design. *Econometrica 73*(6), 1771–1813.

Bergemann, D. and S. Morris (2009a). Robust implementation in direct mechanisms. *The Review of Economic Studies 76*(4), 1175–1204.

Bergemann, D. and S. Morris (2009b). Robust virtual implementation. *Theoretical Economics 4*(1), 45–88.

Bergemann, D. and S. Morris (2011). Robust implementation in general mechanisms. *Games and Economic Behavior 71*(2), 261–281.

Bhaskar, V. and C. Thomas (2018). Community enforcement of trust with bounded memory. *The Review of Economic Studies 86*(3), 1010–1032.

Board, S. and M. Meyer-ter-Vehn (2013). Reputation for quality. *Econometrica 81*(6), 2381–2462.

Brandenburger, A. and E. Dekel (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory 59*(1), 189–198.

Crémer, J. and R. P. McLean (1985). Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica 53*(2), 345–361.

Cripps, M. W., G. J. Mailath, and L. Samuelson (2004). Imperfect monitoring and impermanent reputations. *Econometrica 72*(2), 407–432.

Dekel, E., D. Fudenberg, and S. Morris (2007). Interim correlated rationalizability. *Theoretical Economics 2*(1), 15–40.

Ekmekci, M. (2011). Sustainable reputations with rating systems. *Journal of Economic Theory 146*(2), 479–503.

Eliaz, K. (2002). Fault tolerant implementation. *The Review of Economic Studies 69*(3), 589–610.

Ely, J. C. and M. Peski (2006). Hierarchies of belief and interim rationalizability. *Theoretical Economics 1*(1), 19–65.

Fudenberg, D. and D. K. Levine (1989). Reputation and equilibrium selection in games with a patient player. *Econometrica 57*(4), 759–778.

Fudenberg, D. and D. K. Levine (1992). Maintaining a reputation when strategies are imperfectly observed. *The Review of Economic Studies 59*(3), 561–579.

Gershkov, A., J. K. Goeree, A. Kushnir, B. Moldovanu, and X. Shi (2013). On the equivalence of bayesian and dominant strategy implementation. *Econometrica 81*(1), 197–220.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of law and economics 24*(3), 461–483.

Gul, F. (1998). A comment on aumann's bayesian view. *Econometrica 66*(4), 923–927.

Harsanyi, J. C. (1967). Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science 14*(3), 159–182.

Harsanyi, J. C. (1968a). Games with incomplete information played by "bayesian" players part ii. bayesian equilibrium points. *Management Science 14*(5), 320–334.

Harsanyi, J. C. (1968b). Games with incomplete information played by 'bayesian'players, part iii. the basic probability distribution of the game. *Management Science 14*(7), 486–502.

Hu, T.-W. (2007). On p-rationalizability and approximate common certainty of rationality. *Journal of Economic Theory 136*(1), 379–391.

Jehiel, P., M. Meyer-ter Vehn, and B. Moldovanu (2012). Locally robust implementation and its limits. *Journal of Economic Theory 147*(6), 2439–2452.

Jehiel, P., M. Meyer-ter Vehn, B. Moldovanu, and W. R. Zame (2006). The limits of ex post implementation. *Econometrica 74*(3), 585–610.

Jullien, B. and I.-U. Park (2014). New, like new, or very good? reputation and credibility. *The Review of Economic Studies 81*(4), 1543–1574.

Kajii, A. and S. Morris (1997). The robustness of equilibria to incomplete information. *Econometrica: Journal of the Econometric Society 65*(6), 1283–1309.

Kajii, A. and S. Morris (1998). Payoff continuity in incomplete information games. *journal of economic theory 82*(1), 267–276.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory 27*(2), 245–252.

Lewis, D. (1969). *Convention: A Philosophical Study.* Harvard University Press.

Liu, Q. (2011). Information acquisition and reputation dynamics. *The Review of Economic Studies 78*(4), 1400–1425.

Liu, Q. (2015). Correlation and common priors in games with incomplete information. *Journal of Economic Theory 157*, 49–75.

Lopomo, G., L. Rigotti, and C. Shannon (2009). Uncertainty in mechanism design. *Discussion Paper*.

Marinovic, I., A. Skrzypacz, and F. Varas (2018). Dynamic certification and reputation for quality. *American Economic Journal: Microeconomics 10*(2), 58–82.

Maskin, E. (1992). Auctions and privatization. *Privatization ed. by Horst Siebert*.

Mathevet, L., D. Pearce, and E. Stacchetti (2019). Reputation and information design. *Working paper*.

Mertens, J.-F. and S. Zamir (1985). Formulation of bayesian analysis for games with incomplete information. *International Journal of Game Theory 14*(1), 1–29.

Meyer-ter-Vehn, M. and S. Morris (2011). The robustness of robust implementation. *Journal of Economic Theory 146*(5), 2093 – 2104.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics 12*(2), 380–391.

Monderer, D. and D. Samet (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior 1*(2), 170–190.

Morris, S. (1995). The common prior assumption in economic theory. *Economics & Philosophy 11*(2), 227–253.

Morris, S., R. Rob, and H. S. Shin (1995). p-dominance and belief potential. *Econometrica: Journal of the Econometric Society 63*(1), 145–157.

Myerson, R. B. (1991). *Game theory: analysis of conflict.* Harvard University Press.

Okuno-Fujiwara, M., A. Postlewaite, and K. Suzumura (1990). Strategic information revelation. *The Review of Economic Studies 57*(1), 25–47.

Ollár, M. and A. Penta (2017). Full implementation and belief restrictions. *American Economic Review 107*(8), 2243–77.

Oury, M. and O. Tercieux (2012). Continuous implementation. *Econometrica 80*(4), 1605–1637.

Oyama, D. and O. Tercieux (2010). Robust equilibria under non-common priors. *Journal of Economic Theory 145*(2), 752 – 784.

Rosenthal, R. W. (1979). Sequences of games with varying opponents. *Econometrica: Journal of the Econometric Society 47*(6), 1353–1366.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under" almost common knowledge". *The American Economic Review 79*(3), 385–391.

Tercieux, O. (2006). p-best response set. *Journal of Economic Theory 131*(1), 45–70.

Van Der Schaar, M. and S. Zhang (2015). A dynamic model of certification and reputation. *Economic Theory 58*(3), 509–541.

Weinstein, J. and M. Yildiz (2007). A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica 75*(2), 365–400.

Yamashita, T. (2015). Strategic and structural uncertainty in robust implementation. *Journal of Economic Theory 159*, 267–279.