Essays on Information Economics

Thesis by

Fan Wu

In Partial Fulfillment of the Requirements for the

Degree of

PhD

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2025

Defended April 29, 2025

Fan Wu
ORCID: [0009-0003-8229-6625]

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Omer Tamuz, and my committee chair, Luciano Pomatto, for their continuous support across several research projects. They both taught me how to generate ideas and how to write papers. My first publication in JET benefited significantly from their guidance. They pointed out unclear parts of my writing so I could revise them, and sometimes directly corrected my drafts—from which I learned a great deal. They also taught me how to respond to referee reports. I am profoundly grateful for the immense effort they invested in mentoring me.

In my fourth year, I began working on my job market paper. Although I had a clear motivating story in mind, I struggled to formulate an appropriate model. I considered giving up on the idea several times. It was Omer who encouraged me to keep trying, believing in the idea and offering valuable suggestions on the modeling choices. Later, while searching for a characterization, I initially found only a sufficient condition. I was stuck at that stage for a month, thinking about it for over ten hours a day. I could not prove that the condition was necessary, nor could I construct a counterexample. After twenty days of frustration, I shared my struggle with Omer and Luciano. Omer, sympathetic, said I already had enough for a job market paper. Luciano, on the other hand, insisted I could go further. He told me that constructing counterexamples deepens understanding and is never a waste of time. Eventually, I managed to construct a counterexample, and the next day, I obtained a necessary and sufficient condition. When I excitedly told Luciano, his first reaction was, “No way!” I still wonder whether he had believed I would get there in the first place. :)

My advisor has always been incredibly patient. During discussions about math or proofs, I sometimes lose my temper and raise my voice, but he remains calm and composed. Luciano can be blunt but honest—and often, that was exactly what I need. I recall showing him results that generalized earlier work. His response was, “Why are you showing me this? Why is this useful?” In hindsight, those words pushed me in the right direction. I moved on to a new project that became my second JET publication.

I am also grateful to Yi Xin, now my coauthor, who invited me to work on a project about the insurance industry. She was always optimistic while I was more skeptical. I developed the model and she tested it with data. Time and again, she told me the

model did not perform well and pushed me to refine it. Without her persistence, the project would never have come to fruition. An estimation technique I developed for that paper became the basis for my second job market paper. One of her committee members asked whether we could prove that the technique is a contraction. The problem was daunting, involving a proof of contraction in a functional space. Yi remained confident, while I was overwhelmed. On the 14th day of working on it, I was ready to give up and explained the insurmountable difficulties to her. I vividly remember her being in shock. Still, she asked whether there might be an economic interpretation at the point where I was stuck. That question sparked an idea. Suddenly, the missing pieces came together. I could not stop talking through the proof, even though she had to pick up her daughter—so I explained the argument in her car as she drove to CCC. That evening, I completed almost all major steps of the proof.

I also thank Peter Caradonna, Alexander Bloedel, and Axel Niemeyer for their insightful comments on my papers and help preparing for the job market. Peter gave me countless suggestions on delivering pitches and presentations. Whenever I am uncertain or stuck, he is the first person I turn to. He knows many mathematical results I am unfamiliar with. I first met Alexander when he was a postdoc at Caltech and was struck by his deep knowledge of the literature and his insightful feedback. He has always been generous with his time, even after moving to UCLA. He helped train me for interviews and job talks, offering honest and constructive feedback. Axel is the only faculty member at Caltech specializing in mechanism design, and we share many research interests. I only wish he had joined Caltech earlier. My job market paper benefited greatly from his comments.

I thank Kim Border for his extensive lecture notes. Sadly, he passed away during my second year, and I never had the chance to speak with him in person. Fortunately, many of his notes are preserved in the Kim Border Repository. I read over a thousand pages of his work. He was an excellent educator and conveyed mathematical ideas with clarity and insight.

I also thank Xin Shan for providing the motivating story behind my job market paper.

I am grateful for the companionship and support of the Hardcore Study Group—Kexin Feng and Ke Shi—my closest friends at Caltech. We have supported each other since our second year. Time flies, and I wish you both the best in your careers ahead.

I thank Zhuofang Li for listening to several tentative versions of my job market paper proof. On Christmas Day 2023, as I approached the final stages of the proof, the entire Baxter building was empty and Zhuofang was the only one still in office. Whenever I made progress, I would present to her. Although she does not work in theory, she patiently listened, which helped me organize and solidify my argument. Similarly, I thank Jake Zhang for listening to my second job market paper proof. Peter Doe, the only other theorist in my cohort, provided helpful conversations about my papers and presentations—and was also a great neighbor.

I thank Emily Miao for playing squash with me. She is my best friend outside our department. When I first met her, I casually suggested we form a squash team. I was not serious, but she actually went on to build a team of 80 members. I hope we shall play more in the future.

I thank Mu Zhang, Kun Zhang and Harry Pei for their helpful comments on my research projects.

I thank Raghu C Murthy, a retina specialist, for fixing my retina in my second year.

Finally, I thank Omer Tamuz and Yi Xin for supporting my conference travel.

There are so many people who have helped me during my PhD journey. Please forgive me if I have inadvertently left your name out.

ABSTRACT

This paper on information economics contains three chapters. In the first chapter, we study how to incentivize information acquisition in a principal-agent model. A principal hires an agent to collect information about a state. We study the optimal contract that incentivizes the agent to acquire the most precise information. In the second chapter, we study how to recover information in the selection model. We show that, given the selection rule and the observed selected outcome distribution, the entire outcome distribution can be characterized as the fixed point of an operator, which we prove to be a functional contraction. In the third chapter, we study how to implement randomized allocation rules with outcome-contingent transfers.

PUBLISHED CONTENT AND CONTRIBUTIONS

Liu, Yi and Fan Wu (2024). "Implementing randomized allocation rules with outcome-contingent transfers". In: *Journal of Economic Theory* 220, p. 105878. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2024.105878. URL: https://www.sciencedirect.com/science/article/pii/S002205312400084X. Fan Wu came up with the conceptualization of this work, analyzed and solved the model, and wrote the entire manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INCENTIVIZING INFORMATION ACQUISITION

1.1 Introduction

In many situations, an agent is tasked with collecting information about a quantity of interest. Examples include a statistician collecting data for the Bureau of Statistics, a meteorologist forecasting weather for the National Weather Service, or a consultant assessing the profitability of a new market. A natural question is how to design a contract to incentivize the agent to collect information and report it accurately.

A difficulty in designing such contracts is that the agent's effort and the information acquired are often unobservable to the principal. For example, a statistician might be able to fabricate part of their data. Similarly, in the case of a meteorologist analyzing weather patterns, the amount of time and effort spent on the analysis might not be easily observable. In the case of the consultant, predicting market profitability may rely on private connections or specialized skills, which are often proprietary. Therefore, the contract between the principal and the agent must take into account moral hazard.

In this paper, I study the design of an optimal contract for a principal (she) that incentivizes the agent (he) to gather information and maximize its precision. The agent conducts a location-scale experiment that is centered around the state and controls the precision (the inverse of the scale) at a cost. Eventually, the state is revealed and the contract can depend on the state and the agent's report. I abstract from the principal's broader decision problem—that is, the way in which the principal uses the information generated by the agent—and focus on the problem where the principal's objective is only to maximize precision subject to some fixed budget constraint on transfers. My analysis and results shall be robust to whatever the broader decision problem may be. The budget reflects the maximum amount of transfer that the principal is able to use.

In practice, a budget that can only be used for a specific task is commonly observed. For example, government agencies typically operate within budget limits imposed by government funding and must return any unused funds to the government.[1]

[1] Schick (2008) provides an in-depth look at how the federal budget is structured and constrained by funding limits imposed by the government.

There are numerous reports of unspent government funding being returned to the government, such as Mansouri (2024), Cuccia, 2023.[2] In firms, it is also common for agents to operate under a budget and to eventually return unused funds to the firm. Anthony et al. (2007) and Horngren (2009) provide comprehensive empirical evidence for this practice.[3] In the consulting example, the transfer could take the form of a promotion, a recommendation, or a grade. Sometimes, budgets are restricted by external funding designated for specific purposes, such as university grants.

In such settings, a simple contract is to allocate the entire budget to the agent if their report is close to the actual state, and withhold payment otherwise. I call these incentive schemes *cutoff transfers*. The cutoff transfer is the simplest transfer rule; it is straightforward to understand and easy to implement. The literature on principal-agent problems has long been interested in when simple contracts are optimal (Carroll, 2015; Herweg, Müller, and Weinschenk, 2010; Oyer, 2000; Gottlieb and Moreira, 2022).

My main result is that cutoff transfers are optimal across a large variety of settings. In particular, I identify a sufficient and necessary condition on the agent's information structure such that for all cost functions, there exists an optimal transfer that is a cutoff transfer.

Formally, I study a model in which the principal wants to incentivize the agent to acquire information on an underlying state $\theta \in \mathbb{R}$. For simplicity, I assume the state θ admits an improper uniform common prior.[4] The agent can acquire a costly signal $s \in \mathbb{R}$ about the state. The signal takes the form $s = \theta + \frac{1}{\lambda}\varepsilon$ where ε is drawn from some symmetric and single-peaked probability density function, and λ is a measure of precision controlled by the agent. An important example is the case where ε admits a standard Gaussian distribution. The agent chooses precision λ at a cost $c(\lambda)$.

The principal observes neither the agent's signal s nor the agent's choice of precision λ. The agent makes a report $a \in \mathbb{R}$ after observing the signal. The principal wants

[2]Federal regulations also specify where unspent grant money is returned, as seen at the National Archives, 2024.

[3]Anthony et al. (2007) offer an extensive analysis of organizational budget management, highlighting the frequent practice of returning unspent funds to a central or general fund in both private and public sectors. Horngren (2009) explores cost control and budget allocation, noting how unused resources are often returned to a central pool for future use.

[4]I generalize the model to the case where the state is multi-dimensional and the case with Gaussian prior.

to incentivize the agent to maximize precision and truthfully report his signal. She does so by a transfer rule t that depends on the report a and the state θ, which is eventually observed by both players.[5] I impose limited liability ($t \geq 0$) and limited budget. The budget can only be used for this task. The agent's payoff is his expected utility on transfer minus the cost. I allow for any increasing utility function, which accommodates arbitrary risk attitude of the agent.

I refer to a transfer rule as *optimal* if it induces the maximum precision among all transfers and elicits truthful reports. There are at least two difficulties in this problem. First, the transfer design problem is an infinite-dimensional optimization problem, as the transfer rule itself is a function. Second, there are two layers of incentive compatibility issues. The principal does not observe the precision chosen by the agent, which is a moral hazard problem, and she does not observe the signal either, leading to a communication problem.

I say that a transfer rule is a *cutoff* transfer if it pays the entire budget when the distance between the report and the state is below a cutoff d and pays 0 otherwise. For which signal distributions are cutoff transfers optimal?

My main result (Theorem 1) identifies a sufficient and necessary condition for cutoff transfers to be optimal. This condition is weaker than the monotone likelihood ratio property (Milgrom, 1981; Rogerson, 1985; Jewitt, 1988) and is satisfied by most common distributions, including Gaussian, Laplace, logistic, and the uniform distribution. Theorem 1 shows that if this condition holds, then for all cost functions cutoff transfers are optimal. Conversely, if the condition does not hold, then there is a cost function for which no cutoff transfer is optimal. Moreover, this cost function is not pathological and can be taken to be increasing and differentiable. Furthermore, when the condition holds, I characterize the optimal cutoff (Theorem 2).

An intuition for the optimality of cutoff rules is that they provide the strongest incentives among all contracts. Loosely speaking, since more precise signal structures generate signals that are "more concentrated" around the state, a precision-maximizing contract must pay the agent more for the report closer to the state. The cutoff transfers take this logic to the extreme: the principal exhausts her entire budget when the report is sufficiently close to the state and pays nothing otherwise. Importantly, however, this intuition is incomplete, because the notion of "more con-

[5] I study the case where the state is unobservable but the principal has a private signal about the state later.

centrated" signals is imprecise. My condition is exactly what is needed to make this logic tight.

All the results generalize to the n-dimensional case. The signal distribution is still symmetric, single-peaked, and centered around θ. Here the symmetry means that the density of the distribution depends only on the Euclidean distance from the state θ. A cutoff transfer pays 1 when the Euclidean distance between the report and the state is less than a cutoff d. In this setting I show that an analogous result holds (Proposition 2).

I also generalize my results to the case of a proper Gaussian prior and Gaussian signals. I show that for all cost functions, cutoff transfers are optimal (Proposition 3).

In addition, I extend my results to the setting where the state is unobservable. Instead, the principal privately receives a signal about the state. When designing the transfer, she uses her signal instead of the state to discipline the agent. I show that given a Gaussian or uniform prior and Gaussian signal, cutoff transfers are optimal (Proposition 4).

Lastly, I apply my results to a classic principal-agent problem, offering new insights into the optimality of simple contracts. In the classic setting, a principal incentivizes the agent to produce an output. Unlike the traditional framework, in which the principal's goal is to maximize expected payoff (of outputs) minus transfers, I assume the principal's sole objective is to maximize output, constrained by a budget limit.[6] As a corollary of my main result, the monotone likelihood ratio property is sufficient to ensure the optimality of cutoff transfers (Corollary 3). Here, the cutoff transfer pays the entire budget if the output is above a cutoff. This corollary provides an alternative explanation for simple contracts: the optimality of cutoff transfers stems from the nature of budgets in many contracting scenarios. In contrast, closed-form solutions for the optimal transfers in the classic setting are generally not obtainable (see Bolton and Dewatripont, 2004, Chapter 4.5).

One point to note is that my proof relies on techniques from monotone comparative statics. This allows me to analyze the problem under weak assumptions on the cost function and signal distribution. In particular, I do not need to assume the validity of the first-order approach (Rogerson, 1985; Jewitt, 1988).

[6]This is a common practice in firms; see Anthony et al. (2007) and Horngren (2009).

Related Literature

My paper is related to the literature on incentivizing information acquisition (see, e.g., Osband, 1989; Li and Libgober, 2023; Li, Hartline, et al., 2022; Neyman, Noarov, and Weinberg, 2021; Zermeno, 2011; Carroll, 2019; Chen et al., 2023; Chade and Kovrijnykh, 2016; Whitmeyer and Zhang, 2023; Sharma, Tsakas, and Voorneveld, 2023; Clark and Reggiani, 2021).[7]

In Li, Hartline, et al. (2022), an agent exerts a binary level of effort to refine a posterior from a prior. The paper studies optimizing proper scoring rules by maximizing the increase in the score with effort. In Li and Libgober (2023), a principal hires an agent to learn about a binary state. The agent acquires information over time through a Poisson information arrival technology. The principal rewards the agent with a fixed-value prize as a function of the agent's sequence of reports and the state. Li and Libgober (2023) identify conditions under which it is without loss to elicit a single report after all the information has been acquired.

Osband (1989) studies a principal with a quadratic prediction cost who incentivizes an expert to collect information. The expert can increase the precision (the inverse of variance) of the prediction at a constant cost. The principal minimizes the sum of the expected error variance and the expected transfer. In this stylized setting, the optimal transfer consists of a quadratic report error term plus a linear term on an initial belief error, with three parameters.

In Whitmeyer and Zhang (2023) and Sharma, Tsakas, and Voorneveld (2023), a rational inattentive agent can acquire information flexibly subject to a posterior separable cost. The principal wants to minimize the expected monetary cost of implementing a given information structure. Similarly, Clark and Reggiani (2021) study the Pareto optimal contract that maximizes social welfare. In all these papers, both the experiment and the signal are unobservable to the principal. In Rappoport and Somma (2017), a principal hires an agent to acquire costly information to influence a third party's decision. This paper assumes that the realized piece of information is observable and contractible.

Zermeno (2011) and Papireddygari and Waggoner (2022) study menu design with information acquisition. In their setting, the principal first offers a menu of contracts.

[7] In Neyman, Noarov, and Weinberg (2021), a rational expert aims to predict the probability of a biased coin flip. He acquires information by choosing the number of flip trials at a fixed cost per flip. Neyman, Noarov, and Weinberg (2021) study the optimal scoring rule that incentivizes precision.

Then the agent privately acquire costly information. Next, the agent selects a contract from the menu. The selection therefore reveals some information the agent acquired.

In Carroll (2019), the principal is uncertain about the expert's information acquisition technology and only knows some experiments that the agent can choose. The principal evaluates the incentive contract by a worse-case criterion.

Argenziano, Severinov, and Squintani (2016) and Kreutzkamp (2023) study costly information acquisition and transmission. In both papers, there is no transfer and the agent cares about the principal's action. In Kreutzkamp (2023) setting, the sender publicly chooses an experiment. In Argenziano, Severinov, and Squintani (2016), the expert acquires information by choosing a number of binary trials to perform.

In the classic principal-agent model, several papers also show the optimality of the cutoff transfer but rely on different assumptions. Oyer (2000) assumes the validity of the first-order approach, the existence of the optimal contract, and the absence of the IR constraint. He shows that cutoff transfers are optimal among monotone contracts for a risk neutral agent. Herweg, Müller, and Weinschenk (2010) show that cutoff transfers are optimal for expectation-based loss averse and risk neutral agent.

Notably, there is a growing literature that adopts the same assumption that the principal cannot take money from the agent and can only reward the agent with a prize for which they have no other uses; see Li and Libgober (2023), Li, Hartline, et al. (2022), Deb, Pai, and Said (2018), Deb, Pai, and Said (2023), Dasgupta (2023), Hébert and Zhong (2022), and Wong (2023).

1.2 Model

I study a principal-agent model, where the principal (she) wants to incentivize the agent (he) to acquire information regarding an underlying state $\theta \in \mathbb{R}$. They share a common prior over θ. It will be convenient to assume the prior to be an improper uniform prior on $\mathbb{R}$. I generalize my results to the case in which the state is multi-dimensional in Section 1.5, and I study the case of a proper Gaussian prior in Section 1.5.

The agent can acquire a costly signal $s \in \mathbb{R}$ of the form

$$s = \theta + \frac{1}{\lambda}\varepsilon,$$

where ε is drawn from a distribution with a symmetric and single-peaked probability density function (PDF) ϕ. I assume ϕ is continuously differentiable. The function ϕ

is supported on an interval, which can be bounded or unbounded. The parameter λ is a measure of the signal's precision, a scale parameter that is inversely proportional to the standard deviation of s. The PDF of the signal, given θ and λ, is denoted by $\varphi(\cdot;\theta,\lambda)$. It takes the form

$$\varphi(x;\theta,\lambda) = \lambda\phi(\lambda(x-\theta)).$$

For example, if ϕ is the PDF of the standard Gaussian distribution, then $\varphi(\cdot;\theta,\lambda)$ is the PDF of a Gaussian distribution with mean θ and standard deviation $1/\lambda$. Note that $\phi(x) = \varphi(x;0,1)$.

The agent chooses the precision λ at a cost $c(\lambda)$. For example, suppose the agent's signal is the aggregation of many small independent signals. Then as the number of small signals becomes large, the aggregate signal tends to a Gaussian distribution. If the agent incurs a cost that depends on the number of small signals he gathers, then the agent effectively controls the standard deviation of the aggregate signal at a cost. I assume that the cost function c is lower semicontinuous.[8] The pair (ϕ, c) constitutes the primitive of the model.

The principal observes neither the agent's signal s nor the agent's choice of precision λ. Instead, after observing the signal, the agent sends a report $a \in \mathbb{R}$ to the principal. Eventually both players observe the state θ.[9] The principal wants to incentivize the agent to maximize precision and truthfully report his signal. She does so by means of a transfer. The transfer can depend on the state and the agent's report. The assumption that the realized state is contractible is familiar from the literature on belief elicitation via (proper) scoring rules and prediction markets, ubiquitous in the principal-expert literature, and well-suited to economic applications in which the state is publicly observable ex post (e.g., the outcome of an election being forecast by a pollster, or the conditions of a new market being analyzed by a consultant).

I assume that the transfer t is a function of the difference $\theta - a$ and it vanishes at infinity. That is, it satisfies

$$\lim_{x\to-\infty} t(x) = \lim_{x\to\infty} t(x) = 0.$$

As I discuss below, the assumption that the transfer depends only on the difference between $a - \theta$ is to ensure the agent's payoff is well-defined. I also assume limited

[8] It would be natural to also assume that the cost function is increasing—that is, the higher the precision, the higher the cost. Imposing this assumption or not does not affect my results.

[9] I relax this assumption in Section 1.5 where instead of eventually observing the state, they eventually observe a signal about the state.

liability, i.e., that the principal cannot take money from the agent, so that $t \geq 0$. In addition, the principal has a limited budget that does not depend on the state. She can use this budget only for the transfer. Without loss, I take the budget to be 1.

The timing of moves is as follows:

1. The principal commits to a transfer rule $t\colon \mathbb{R} \to [0,1]$.
2. The agent chooses to accept or not. The game continues if the agent accepts and terminates otherwise.
3. The agent privately chooses a signal precision λ.
4. The agent privately observes a signal realization s drawn from the distribution $\varphi(\cdot;\theta,\lambda)$.
5. The agent makes a report $a \in \mathbb{R}$.
6. Finally, the state θ is revealed and the agent receives the transfer $t(\theta - a)$.

The agent's payoff is given by the expected utility of the transfer minus the cost of acquiring information:

$$\mathbb{E}u(t) - c(\lambda),$$

where $u\colon [0,1] \to \mathbb{R}$ is strictly increasing and continuous, and normalized so that $u(0) = 0$, and $u(1) = 1$.[10] To ease the notation, it will be convenient to treat the transfer rule as paying in utils, rather than money. Under my assumptions on u, this is equivalent, and without loss of generality, to assuming u is the identity function. Finally, I assume the agent has an outside option which gives payoff 0 if he rejects the contract.

After choosing λ, the agent observes a signal. Conditional on a signal realization s, it follows from Bayes' rule (adapted to a uniform prior) that the agent's posterior has PDF $\varphi(\cdot;s,\lambda)$. The agent now chooses a report a to maximize the expected transfer, which is therefore given by

$$E(\lambda;t) = \max_{a\in\mathbb{R}} \int_{\mathbb{R}} t(\theta' - a)\varphi(\theta';s,\lambda)d\theta'.$$

[10] Here I can drop the continuity assumption on u and only require u to be weakly increasing. All my results still hold.

Maximizing over report a shows that when choosing the precision, the agent anticipates how he optimally reports later. Thus, the model allows for double deviation. The assumption that t vanishes at infinity ensures that the maximum is well defined.[11]

Note that as t depends only on the difference $\theta - a$, the expected transfer computed above does not depend on s.[12] Consequently, the agent's (interim) expected transfer conditional on signal s does not depend on s, which must coincide with his unconditional expected transfer. If t depends arbitrarily on a and θ, the agent's unconditional expected transfer might not be well-defined, since the distribution of s is improper. In Appendix 1.7, I show that when the agent's expected transfer is well-defined, it is without loss to assume that the transfer depends only on the difference $a - \theta$. An interesting fact is that almost all popular key performance indicators in the forecast industry depend only on the prediction error $a - \theta$ (Vandeput, 2021).

The principal's objective is to maximize λ under the constraint of inducing a truthful report from the agent. (As I show later, this constraint will not be binding.) This objective can be interpreted as capturing, in reduced-form, settings in which the principal uses the information conveyed by the agent's report to solve some (unmodeled) decision problem. Indeed, under broad conditions, it is optimal for the principal to maximize precision regardless of the decision problem that she faces. First, when the density ϕ is strongly unimodal, for a large class of monotone decision problems (Karlin and Rubin, 1956), higher precision is always better for the principal (see Lehmann, 2011, Theorem 5.1 and 5.2).[13,14] Second, when the distribution ϕ is *self-decomposable*, the signal with different precisions are ranked in the Blackwell order; thus the principal is better off with higher precision for all decision problems. The self-decomposable distributions include all stable distributions (such as Gaussian and Cauchy distributions) and some non-stable ones

[11] I prove this in Lemma 5 in the appendix.

[12] To see this, note that

$$\max_a \int_{\mathbb{R}} t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' = \max_a \int_{\mathbb{R}} t(\theta' - a + s)\varphi(\theta'; 0, \lambda)d\theta' = \max_a \int_{\mathbb{R}} t(\theta' - a)\varphi(\theta'; 0, \lambda)d\theta'$$

[13] ϕ is strongly unimodal if $-\ln\phi$ is convex. This is slightly stronger than global increasing elasticity.

[14] The class of *monotone* decision problems is defined in terms of the action space and the permissible loss functions. For each θ, there is a correct action $A(\theta)$. The function $A(\theta)$ is real-valued and nondecreasing. The range of $A(\theta)$ is the action space. The loss function is minimized at the correct action and is nondecreasing as the action moves away from the correct action on either side.

such as Laplace.[15]

1.3 Preliminary Analysis

First, I define a relaxed problem, in which the principal designs a transfer to maximize the agent's signal precision, without incentivizing truthful reporting.

$$\begin{aligned} \underset{\lambda, t}{\text{maximize}} \quad & \lambda \\ \text{subject to} \quad & 0 \leq t \leq 1 \\ & \lambda \in \arg\max E(\cdot; t) - c(\cdot) \quad \text{IC} \\ & E(\lambda; t) - c(\lambda) \geq 0 \quad \text{IR} \end{aligned} \tag{1.1}$$

The first constraint in the relaxed problem is limited liability and limited budget. The second constraint is an incentive compatibility constraint, since the precision λ is chosen by the agent and is unobserved by the principal and cannot be contracted upon.[16] The last constraint is a participation constraint, also known as individual rationality, and is implied by the outside option available to the agent. A transfer rule t is *optimal* if it solves problem (1.1) and induces the agent to report truthfully.

A quantity that will be key to the analysis is the elasticity of the standardized signal distribution ϕ.

Definition 1. The *elasticity* η of signal distribution ϕ at $x > 0$ is defined as

$$\eta(x) = -\frac{d\phi(x)/\phi(x)}{dx/x}.$$

For x such that $\phi(x) = 0$, I let $\eta(x) = +\infty$.

Elasticity measures by how much a percentage change in x leads to a percentage change in the density function $\phi(x)$. For example, if ϕ is the PDF of a standard Gaussian, then $\eta(x) = x^2$. If ϕ is the PDF of a standard Laplace, then $\eta(x) = x$. Notice that both examples have weakly increasing elasticity. In fact, weakly increasing elasticity is satisfied by most common nonatomic distributions defined on an interval, including the uniform distribution, the triangular distribution, and the logistic distribution. The next lemma shows that the weakly increasing elasticity condition is equivalent to a monotone likelihood property: for any $0 \leq x_1 \leq x_2$, the ratio $\frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)}$ increases in λ. That is, a more accurate signal (x_1) is more likely to appear than a less accurate signal (x_2) as precision increases.

[15] See Goel and DeGroot, 1979 and Lehmann, 2011 for a detailed discussion.

[16] The continuity of $E(\cdot; t)$ and the lower semicontinuity of c ensures that the arg max in IC is well defined.

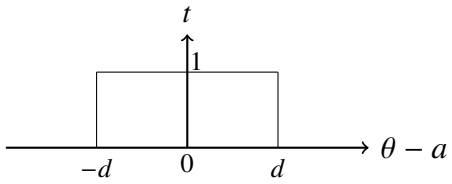


Figure 1.1: Cutoff Transfer.

Lemma 1.

$$\frac{\partial[\ln \phi(\lambda x)]}{\partial \ln \lambda} = -\eta(\lambda x).$$

In particular, for any $0 \leq x_1 \leq x_2$,

$$\frac{\varphi(x_1; 0, \lambda)}{\varphi(x_2; 0, \lambda)} \textit{ increases in } \lambda \quad \Leftrightarrow \quad \eta(\lambda x_1) \leq \eta(\lambda x_2).$$

Next, I define a weaker condition.

Definition 2. The signal distribution satisfies *increasing elasticity above 1* if η single-crosses 1 from below and is weakly increasing after the cross, i.e., if for every $x > 0$

$$\eta(x) > 1 \quad \text{implies} \quad \eta(y) \geq \eta(x) \text{ for all } y > x.$$

I define a useful quantity.

$$\eta^{-1}(1) = \inf\{x \in \mathbb{R}_+ | \eta(x) > 1\}.$$

Note that $\eta^{-1}(1)$ is well-defined, as the density function ϕ is integrable.[17] Later on, when I generalize the state and the signal to be n dimensional, $\eta^{-1}(n)$ is defined similarly.

1.4 Characterization of the Optimal Transfer

A simple transfer rule is the *cutoff transfer* that pays the agent 1 when the distance between the report and the state is less than a cutoff d and pays 0 otherwise (see Figure 1.1). As cutoff transfers are symmetric and single-peaked, they have the additional desirable property of inducing truthful reports by the agent.

As transfer rules are functions, the principal faces an infinite dimensional optimization problem. In my main result I show that when the standardized distribution ϕ

[17] Notice that function $1/x$ is not integrable over any neighborhood around 0. Thus, $\eta(0^+) < 1$ since ϕ is integrable over any neighborhood around 0. Moreover, function $1/x$ is not integrable over any neighborhood around $+\infty$. Thus, $\eta(x) > 1$ must hold for some $x \in \mathbb{R}_+$.

satisfies increasing elasticity above one—which holds for all commonly used signal distributions—the principal chooses a cutoff transfer rule, reducing the problem to a one-dimensional one. Moreover, I show that this property of ϕ is also necessary for cutoff transfer rules to always be optimal.

Theorem 1. *The following are equivalent.*

1. *For all cost functions, there exists an optimal transfer that is a cutoff transfer.*
2. *For all increasing and differentiable cost functions, there exists an optimal transfer that is a cutoff transfer.*
3. *Signal density function ϕ satisfies increasing elasticity above* 1.

Note that Theorem 1 holds even if we assume that all cost functions are differentiable and increasing; that is, the result is not driven by considering ill behaved cost functions.

Once we verify the condition in statement (3), then we do not need to worry about the infinite dimensional optimization problem, as cutoff transfers shall be optimal. Moreover, the condition is very easy to check. We only need to compute the elasticity and check if it is monotone in some region. A direct implication of Theorem 1 is that for all cost functions, there exists an optimal transfer that is a cutoff transfer when the signal distribution is Gaussian, since the elasticity of a Gaussian distribution is increasing. The lower bound 1 in the statement is the dimension of the problem. So far both the state and the report are one-dimensional. In Section 1.5, where I generalize this result to n-dimensional signal and states, the dimension n will replace 1.

Note that there is a unique solution—which must be a cutoff transfer—if we additionally impose three mild assumptions.[18]

1. Signal density function ϕ satisfies strictly increasing elasticity above[19] 1.
2. The cost function is continuously differentiable.
3. The optimal precision is an interior solution.

[18] I prove this uniqueness result at the end of the Proof of Theorem 1.

[19] Density function ϕ satisfies *strictly increasing elasticity above 1* if $\eta(\cdot)$ single-crosses 1 from below, $\{x|\eta(x) = 1\}$ is a singleton, and is strictly increasing after the cross.

In the following, I shall first illustrate the sufficiency part (i.e., why (3) implies (1)). Second, I optimize over cutoff transfers and characterize the optimal cutoff. Third, I provide intuition for the necessity part (i.e., why (2) implies (3)).

I define the agent's expected payoff given transfer t as

$$\pi(t) = \max_{\lambda} E(\lambda; t) - c(\lambda)$$

and the agent's choice of precision given t

$$\boldsymbol{\lambda}(t) = \begin{cases} \max[\arg\max_{\lambda} E(\lambda; t) - c(\lambda)] & \text{if } \pi(t) \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

When analyzing cutoff transfers, with slight abuse of notations, I use $E(\lambda; d)$ to represent the expected transfer at precision λ given the cutoff transfer with cutoff d, i.e., $E(\lambda; d) = E(\lambda; \mathbf{1}_{|\theta - a| \leq d})$. Similarly, I denote by $\pi(d)$ and $\boldsymbol{\lambda}(d)$ the expected payoff and choice of precision for the cutoff transfer d, respectively.

Cutoff Transfers Are Optimal

I explain why increasing elasticity above 1 implies that for all cost functions, cutoff transfers are optimal. Suppose that increasing elasticity above 1 holds. I show that for any transfer rule t, there exists a cutoff transfer that induces a weakly larger precision and truthful report.

Here, for ease of exposition, I assume that η is weakly increasing and consider a transfer rule t that is symmetric in $\theta - a$ and weakly decreases in $|\theta - a|$. Given such a transfer rule, the agent reports truthfully and chooses the precision $\boldsymbol{\lambda}(t)$. I construct a cutoff transfer by choosing a cutoff d such that

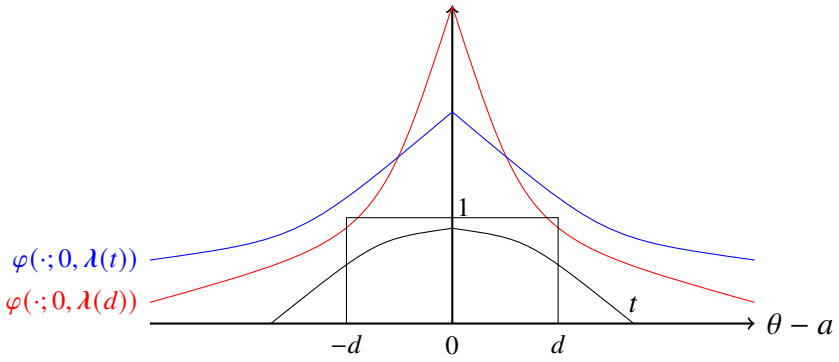
$$E(\boldsymbol{\lambda}(t); d) = E(\boldsymbol{\lambda}(t); t). \tag{1.2}$$

Such a d exists because $E(\boldsymbol{\lambda}(t); d)$ increases in d from 0 to 1 (as the budget is 1). See Figure 1.2 for an example.

Next, I argue that

$$\begin{aligned} &E(\lambda; d) - E(\lambda; t) \geq 0 \text{ for } \lambda > \boldsymbol{\lambda}(t) \\ &E(\lambda; d) - E(\lambda; t) \leq 0 \text{ for } \lambda < \boldsymbol{\lambda}(t). \end{aligned} \tag{1.3}$$

This is depicted in Figure 1.3.



Note: The given transfer t and the matching cutoff transfer d are shown in black. The wider distribution shown in blue is the signal distribution chosen by the agent under t, while the narrower (more precise) distribution in red is the one chosen under d.

Figure 1.2: The Transfer t and The Cutoff Transfer d.

To see this, note that as the cutoff transfer pays the entire budget 1 when $|\theta - a| \leq d$, it is larger than the transfer t within the cutoff region. In addition, Lemma 1 shows that, since the elasticity of ϕ is weakly increasing, for any $0 < x_1 < d < x_2$ the ratio

$$\frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)}$$

increases in λ. Since $E(\lambda;t)$ is the integral of the transfer with respect to the signal distribution, (1.3) follows. As a result, we have $\boldsymbol{\lambda}(d) \geq \boldsymbol{\lambda}(t)$ (see Figure 1.3).

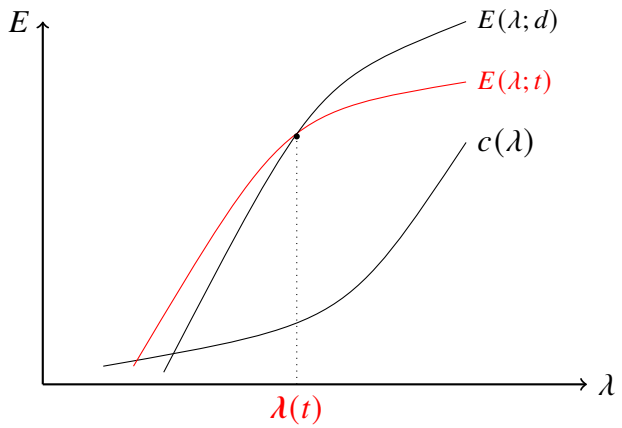


Figure 1.3: Expected Transfer.

In the argument above, I use a stronger condition, global increasing elasticity, to show that cutoff transfers are optimal. By Lemma 1, global increasing elasticity is equivalent to the monotone ratio likelihood property:

$$\text{For all } 0 < x_1 < x_2, \quad \frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)} \text{ is increasing in } \lambda.$$

Thus, this property is a sufficient condition for cutoff transfers to be optimal.

Corollary 1. *If for all $x_2 > x_1 > 0$, $\frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)}$ increases in λ, there exists an optimal transfer that is a cutoff transfer.*

To complete the proof of sufficiency in the general setting, there are two more technical issues. First, for a general transfer t (which is not necessarily symmetric), the agent's report depends on both the signal and the precision. But in our previous example (Figure 1.3), the agent's report equals the signal and does not depend on the precision. Second, the proof sketch above assumed globally increasing elasticity, which is stronger than increasing elasticity above 1. Under this weaker assumption, we do not have for all $x_2 > x_1 > 0$, $\frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)}$ increases in λ. The proof in the appendix deals with both issues.

Optimization of Cutoff Transfers

When ϕ satisfies increasing elasticity above 1, Theorem 1 reduces the infinite dimensional optimization problem to a one-dimensional problem of finding the value of the optimal cutoff. In this section, I will solve this one-dimensional problem and identify the optimal cutoff.

As the cutoff d increases, the agent's expected transfer $E(\cdot; d)$ increases. Let $\bar{d}$ denote the minimum cutoff such that IR constraint holds, i.e.,

$$\bar{d} = \min\{d \geq 0 | \pi(d) \geq 0\}.$$

Notice that $\bar{d}$ depends on the cost function c. The continuity of $E(\cdot; d)$ for all d and the lower semicontinuity of c ensures that this minimum is well defined.

Theorem 2. *If ϕ satisfies increasing elasticity above* 1*, the cutoff transfer*

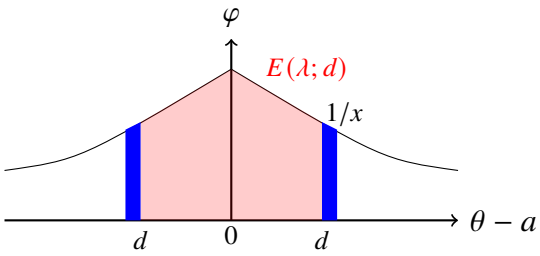
$$d^* = \min\{d \geq \bar{d} | \lambda(d)d \geq \eta^{-1}(1)\}$$

is optimal.

To obtain this result, the following lemma is crucial.

Lemma 2 (Complements or Substitutes)**.** *The expected transfer $E(\lambda; d)$ satisfies:*

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \geq 0 \quad \Leftrightarrow \quad \eta(\lambda d) \leq 1.$$



Note: The agent slightly increases precision from λ to $\lambda + \Delta\lambda$. The area of the red region is the expected transfer $E(\lambda; d)$. The area of two blue regions is the increment of probability that the signal falls into the cutoff.

Figure 1.4: Increment of Expected Transfer When Increasing λ.

This lemma characterizes whether the precision and the cutoff are complements or substitutes. When the cross derivative $\frac{\partial^2 E(\lambda;d)}{\partial\lambda\partial d}$ is positive, increasing the cutoff would increase the marginal return to the precision; thus they are complements. On the other hand, when the cross derivative is negative, the cutoff and the precision are substitutes.

I illustrate the intuition of the boundary case when $\eta(\lambda d) = 1$ and show that

$$\frac{\partial^2 E(\lambda; d)}{\partial\lambda\partial d} = 0.$$

I plot the agent's the expected transfer in Figure 1.4. Since the agent receives transfer 1 when the signal lies in the cutoff, the expected transfer $E(\lambda; d)$ is the probability that the signal lies in the cutoff, which is the area of the light red region. Suppose that the agent slightly increases precision from λ to $\lambda + \Delta\lambda$. The signal shall lie in the cutoff with a higher probability. The area of dark blue region is the incremental probability that the signal lies in the cutoff. Since $\eta(\lambda d) = 1$, in a neighborhood of d, $\varphi(\cdot; 0, \lambda)$ behaves like x^{-1}. We can compute the incremental expected transfer, which coincides with the incremental probability, by the area of the dark blue region

$$E(\lambda + \Delta\lambda; d) - E(\lambda; d) \approx 2\varphi(d; 0, \lambda)\left(d - d\frac{\lambda}{\lambda + \Delta\lambda}\right)$$

which is independent of d as $\varphi(\cdot; 0, \lambda)$ decays at x^{-1}. As a result, $\frac{\partial^2 E(\lambda;d)}{\partial\lambda\partial d} = 0$. A similar argument shows that $\frac{\partial^2 E(\lambda;d)}{\partial\lambda\partial d}$ and $\eta(\lambda d) - 1$ have the opposite sign. Here 1 appears as this problem is one dimensional, which is also why "increasing elasticity above 1" appears in Theorem 1 and $\eta^{-1}(1)$ appears in Theorem 2. Later on when I generalize this problem to n-dimensional, n shows up in both characterizations.

As the expected transfer $E(\lambda; d)$ is the probability that a signal falls within the cutoff region, we have $E(\lambda; d) = 2\Phi(\lambda d) - 1$ where Φ is the CDF of the distribution ϕ.

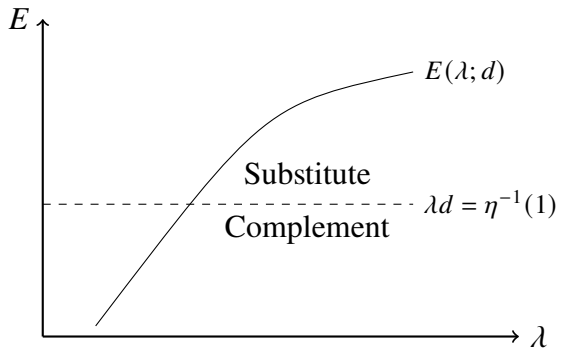


Figure 1.5: Expected Transfer

Consequently, $E(\lambda; d)$ is increasing in λd. Since $\frac{\partial^2 E(\lambda;d)}{\partial \lambda \partial d}$ and $\eta(\lambda d) - 1$ have the opposite sign, by the definition of $\eta^{-1}(1)$, we have

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \geq 0 \quad \text{if} \quad \lambda d \leq \eta^{-1}(1)$$

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \leq 0 \quad \text{if} \quad \lambda d \geq \eta^{-1}(1).$$

Thus, the graph of expected transfer is separated into two parts (see Figure 1.5). When $\lambda d \geq \eta^{-1}(1)$, $\frac{\partial^2 E(\lambda;d)}{\partial \lambda \partial d} \leq 0$, I refer to this part as the *substitute region*. When $\lambda d < \eta^{-1}(1)$, $\frac{\partial^2 E(\lambda;d)}{\partial \lambda \partial d} > 0$, I refer to this part as the *complement region*.

Recall that $\bar{d}$ denotes the minimum cutoff that the agent is willing to work. If $(\bar{d}, \boldsymbol{\lambda}(\bar{d}))$ lies in the substitute region, increasing d shall reduce $\boldsymbol{\lambda}(d)$, which is not optimal for the principal. Consequently, she should set the cutoff at the minimum $\bar{d}$. In this case, the IR constraint is binding and the agent's surplus is zero. If $(\bar{d}, \boldsymbol{\lambda}(\bar{d}))$ lies in the complement region, increasing d shall first increase $\boldsymbol{\lambda}(d)$, and $\boldsymbol{\lambda}(d)d$ shall increase until it hits the boundary $\eta^{-1}(1)$. After hitting the boundary, if we further increase d, then $\boldsymbol{\lambda}(d)$ shall decrease as $(d, \boldsymbol{\lambda}(d))$ enters the substitute region. Thus, the optimal cutoff is the d where $\boldsymbol{\lambda}(d)d$ first hits the boundary $\eta^{-1}(1)$. In this case, the IR constraint is relaxed and the agent enjoys some surplus.

Note that although λ^* is unique by design, the optimal cutoff transfer is not necessarily unique. To see this, recall that changing d causes the curve $E(\cdot; d)$ to rotate. If the cost function c has a kink and $\boldsymbol{\lambda}(d)$ is stuck at this kink, rotating $E(\cdot; d)$ slightly might not affect $\boldsymbol{\lambda}(d)$, which gives rise to multiple optimal cutoff transfers. The cutoff d^* in Theorem 2 is the optimal cutoff transfer that provides the strongest local incentive around λ^*, i.e., it has the largest derivative $\frac{\partial E(\lambda;d)}{\partial \lambda}\big|_{\lambda=\lambda^*}$ among all optimal cutoff transfers. But this example is not generic due to the kink in the cost function.

This is also why a continuous differentiable cost function helps us to get a unique optimal transfer rule.

The Necessity of Increasing Elasticity above 1

In this section, I explain the intuition behind the necessity result—(2) implies (3)—in Theorem 1. That is, why increasing elasticity above 1 is necessary for a cutoff transfer to always be optimal.

I first provide an example where ϕ does not satisfy increasing elasticity above 1. In this case, for some increasing and differentiable cost function, all cutoff transfers are suboptimal.

Example 1. The standardized signal distribution is truncated $\exp(1/x)$:

$$\phi(x) = \begin{cases} k\exp(1/\epsilon), & \text{if } x \in [0, \epsilon), \\ k\exp(1/x), & \text{if } x \in [\epsilon, 1], \\ 0, & \text{otherwise,} \end{cases}$$

where k is a normalizing factor (see the left panel of Figure 1.6). I plot the elasticity of ϕ on the right panel of Figure 1.6. The signal distribution does not satisfy increasing elasticity above 1 as $\eta(x)$ is decreasing for $x \in [\epsilon, 1]$. Fix a pair (λ^*, d^*) such that $\epsilon < \lambda^* d^* < 1$. Suppose an increasing cost function $c \in C^1$ is tangent to $E(\cdot; d^*)$ at $\lambda = \lambda^*$ and $c(\lambda)$ is strictly above $E(\lambda; d^*)$ for all $\lambda \neq \lambda^*$ (see the right panel of Figure 1.7).

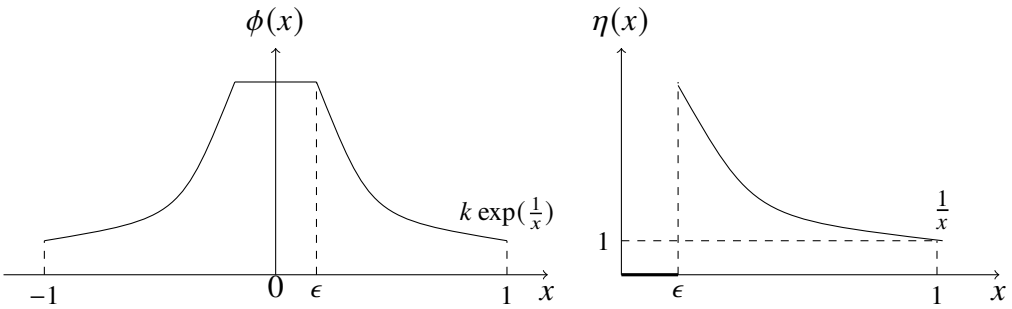


Figure 1.6: Signal Distribution and Elasticity

First, note that d^* is the best cutoff among all cutoff transfers: by design, d^* is the minimum cutoff for the agent to work. By Lemma 2, as (λ^*, d^*) lies in the substitute region, any cutoff transfer with cutoff $d > d^*$ induces a smaller precision. Thus, d^* is the best cutoff transfer.

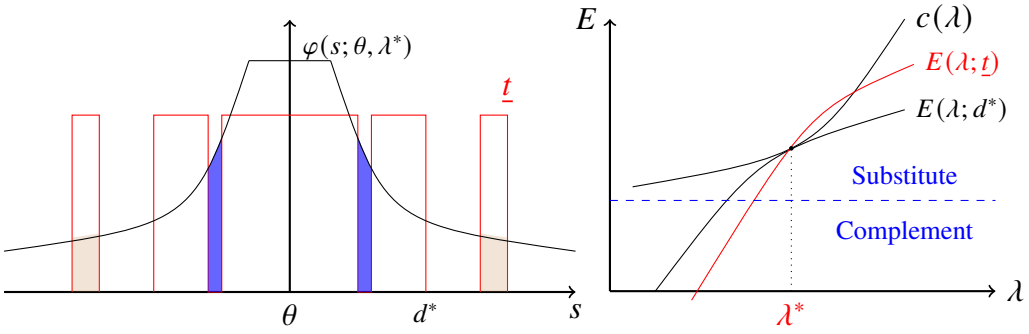


Figure 1.7: The New Transfer Rule and The Expected Transfer

Second, I construct a new transfer rule $\underline{t}$ which induces a strictly larger precision. Starting with the best cutoff transfer rule d^*, I modify the transfer rule by setting one interior region (dark blue) within the cutoff to 0 and one exterior region (light brown) outside the cutoff to 1 (see the left panel of Figure 1.7). These two regions are chosen such that the area of the dark blue region and the light brown region are the same. Note that the area of each region is the expected transfer contributed by this region at precision λ^*. Thus, the new transfer rule $\underline{t}$ has the same expected transfer as cutoff transfer d^* at precision λ^*.

As the signal distribution features decreasing elasticity, the density function $\varphi(s;\theta,\lambda^*)$ is steeper at the blue region. Once we increase precision, the blue area shall shrink more than the brown area. This implies that the new transfer rule $\underline{t}$ has higher expected transfer than the cutoff transfer d^* for slightly higher precision. I translate this comparison to the right panel of Figure 1.7. For higher precision $\lambda > \lambda^*$, $E(\lambda;\underline{t}) > E(\lambda;d^*)$. Thus, $\underline{t}$ shall induce a higher precision.

This example is key to understand more generally the necessity of increasing elasticity above 1. I call a pair (λ^*, d^*) *exposed* if the cutoff transfer d^* is the optimal transfer for some increasing cost function $c \in C^1$ and induces precision λ^*. A pair (λ^*, d^*) being exposed implies that $\eta(x_1) \leq \eta(x_2)$ for all $0 \leq x_1 \leq \lambda^* d^* \leq x_2$. Otherwise, I can use the construction in Example 1 to construct a new transfer than induces a strictly larger precision. In Theorem 2, all (d^*, λ^*) pairs in the substitute region are exposed, as I can always find an increasing and continuously differentiable cost function that is tangent to $E(\cdot;d^*)$ at λ^* and strictly above $E(\cdot;d^*)$ everywhere else. Similarly, when for all increasing and differentiable cost functions there exists an optimal transfer that is a cutoff transfer, all pairs with $\eta(d^*\lambda^*) \geq 1$ are exposed. This implies increasing elasticity above 1. However, as the construction in the ex-

ample above only works in the substitute region, pairs in the complement regions are not exposed. Consequently, we do not have monotonicity for η when it is below 1.

Comparative Statics

Next I study how the optimal cutoff changes with the cost function. Without loss of generality, I assume that the cost function is weakly increasing.[20] I express dependence on the cost function. Let $d^*(c)$ and $\lambda^*(c)$ denote the optimal cutoff and precision given cost function c. Consider two cost functions $c_1 \leq c_2$ where c_1 is less costly. Intuitively, if the agent's cost is higher, the optimal cutoff increases. The next result formalizes this intuition.

Proposition 1 (Comparative Statics)**.** *Suppose that ϕ satisfies increasing elasticity above 1 and $c_1 \leq c_2$. If $c_2(\lambda) - c_1(\lambda)$ is weakly increasing in λ, then $d^*(c_1) \leq d^*(c_2)$ and $\lambda^*(c_2) \leq \lambda^*(c_1)$. In particular, if $c_2 = kc_1$ for some constant $k > 1$, then $d^*(c_1) \leq d^*(c_2)$ and $\lambda^*(c_2) \leq \lambda^*(c_1)$.*

The cost difference $c_2(\lambda) - c_1(\lambda)$ being weakly increasing occurs if c_2 is more convex than c_1. Another interesting comparative statics is to change the budget from 1 to $1/k$. This is equivalent to keeping the budget at 1 and change the cost function from c_1 to kc_1. Thus, lowering the budget would lead to a larger optimal cutoff.

As a direct corollary, I can study what happens if the agent's signal is more precise. Suppose that $\varepsilon_2 = k\varepsilon_1$ with some constant $k > 1$. Recall that the agent's signal is $s = \theta + \frac{1}{\lambda}\varepsilon$. Thus, a smaller noise ε_1 corresponds to a more precise signal, which leads to a smaller cutoff by the next result.

Corollary 2. *Suppose that the cost function is weakly convex. If $\varepsilon_2 = k\varepsilon_1$ with some constant $k > 1$, $d_1^* \leq d_2^*$.*

What if the cost difference $c_2 - c_1$ is not increasing? Then the answer is more involved. Denote by $\boldsymbol{\lambda}(d; c)$ the induced precision given cutoff transfer d and the cost function c. Given the cost function c_1, the minimum cutoff $\bar{d}(c_1)$ and the induced precision $\boldsymbol{\lambda}(\bar{d}(c_1); c_1)$ pair lies either in the substitute region $\boldsymbol{\lambda}(\bar{d}(c_1); c_1)\bar{d}(c_1) \geq \eta^{-1}(1)$ or not. If this pair lies in the substitute region, then the optimal cutoff

[20] Given any cost function $\tilde{c}$ that is not weakly increasing, I can define a new cost function c to be the largest weakly increasing function that is below $\tilde{c}$. The agent's precision choice problem under cost function c is the same as under cost $\tilde{c}$, because $E(\lambda; d)$ is weakly increasing in λ.

coincides with the minimum cutoff, $d^*(c_1) = \bar{d}(c_1)$, by Theorem 2. Then given a larger cost function c_2, the minimum cutoff must be larger, $\bar{d}(c_2) \geq \bar{d}(c_1)$, which entails a larger optimal cutoff, $d^*(c_2) \geq d^*(c_1)$.

However, if the $\boldsymbol{\lambda}(\bar{d}(c_1); c_1)$, $\bar{d}(c_1)$ pair lies in the complement region, the comparison of d^* can go both directions. If $c_2 - c_1$ is decreasing, it may render higher precision relatively more attractive to the agent under cost c_2 than c_1. This may induce a higher precision for each cutoff transfer. As the optimal precision and optimal cutoff are substitutes by Theorem 2, a higher induced precision leads to a lower optimal cutoff.

Readers may wonder what happens to the optimal precision λ^* given $c_2 \geq c_1$. It turns out that increasing the cost function can shift λ^* in both directions, depending on the shape of the cost difference $c_2 - c_1$.

1.5 Extensions

The Multi-Dimensional Case

My results can generalize to the case with a multi-dimensional state. In this section, suppose the state, the signal, and the report are n-dimensional, i.e., $\theta,\ s,\ a \in \mathbb{R}^n$. The signal distribution is still symmetric, single-peaked, and centered around θ. Here symmetry means that the density of the distribution depends only on the Euclidean distance from the state θ. A cutoff transfer pays 1 when the Euclidean distance between the report and the state is less than a cutoff d. The next proposition shows that a statement analogous to that of Theorem 1 applies in this case too.

Proposition 2. *Suppose the state is n-dimensional. The following statements are equivalent.*

1. *For all cost functions, there exists an optimal transfer that is a cutoff transfer.*
2. *For all increasing and differentiable cost functions, there exists an optimal transfer that is a cutoff transfer.*
3. *Density function ϕ satisfies increasing elasticity above n.*

Moreover, if ϕ satisfies increasing elasticity above n, the cutoff transfer

$$d^* = \min\{d \geq \bar{d} | \boldsymbol{\lambda}(d) d \geq \eta^{-1}(n)\}$$

is optimal.

Note that the increasing elasticity condition of Theorem 1 is generalized in this proposition to increasing elasticity above the dimension n, providing an explanation for why the number 1 appeared in Theorem 1.

Gaussian Prior and Gaussian Signal

In this section, I assume that both the prior and the signal admit Gaussian distributions. Since the prior is no longer uniform, a slight adjustment is that now the principal wants the agent to truthfully report the posterior mean rather than the signal.

Let the prior distribution be $\mathcal{N}(0, 1/\lambda_0^2)$. Conditional on a signal s, the agent's posterior is also Gaussian

$$\mathcal{N}(\frac{s\lambda^2}{\lambda_0^2+\lambda^2}, \frac{1}{\lambda_0^2+\lambda^2}),$$

where λ is the agent's choice of precision. I let Λ denote the precision of the posterior

$$\Lambda(\lambda) = \sqrt{\lambda_0^2+\lambda^2}.$$

Then, the expected transfer becomes $E(\Lambda(\lambda); t)$ instead of $E(\lambda; t)$. Therefore, $\Lambda(\lambda)$ shall replace λ in the characterization. Other than this difference, the results remain the same as in previous sections, as the posterior is Gaussian whose elasticity is increasing.

Proposition 3. *Suppose the prior and the signal admit Gaussian distribution. There exists an optimal transfer that is a cutoff transfer. Moreover, the cutoff transfer*

$$d^* = \min\{d \geq \bar{d} | \Lambda(\boldsymbol{\lambda}(d))d \geq \eta^{-1}(1)\}$$

is optimal.

Unobserved State

In this section, I extend my results to the setting where the state is unobservable. Instead, the principal has some private information about the state. When designing the transfer, she uses her private information instead of the state to discipline the agent.[21]

[21]Note that the principal does not have utility for their budget, and so have no incentive to lie about their private signals. Thus, it is without loss of generality to assume that the principal has commitment power.

I assume that the prior is uniform or Gaussian $\mathcal{N}(0, 1/\lambda_0^2)$. The principal wants the agent to truthfully report his posterior mean and maximizes λ. The principal and the agent receive independent Gaussian signals: $s_p \sim \mathcal{N}(\theta, 1/\lambda_p^2)$ and $s \sim \mathcal{N}(\theta, 1/\lambda^2)$. I show that it is without loss to focus on cutoff transfers that pay 1 when $|s_p - a|$ is less than a cutoff and 0 otherwise.

Proposition 4. *Suppose that the state is unobservable and the prior admits uniform or Gaussian distribution. The principal and the agent receive independent Gaussian signals. There exists an optimal transfer that is a cutoff transfer. Moreover, the cutoff transfer*

$$d^* = \min\{d \geq \bar{d} | \Lambda(d)d \geq \eta^{-1}(1)\}$$

is optimal where $\Lambda(d) = (\frac{1}{\lambda_p^2} + \frac{1}{\lambda^2(d)})^{-\frac{1}{2}}$ *for the uniform prior and* $\Lambda(d) = (\frac{1}{\lambda_p^2} + \frac{1}{\lambda^2(d)+\lambda_0^2})^{-\frac{1}{2}}$ *for the Gaussian prior.*

1.6 Implications for Classic Principal-Agent Problem

In this section, I demonstrate how my results apply to a classic principal-agent problem, offering new insights into the optimality of simple contracts. In the classic setting, a principal incentivizes the agent to produce an output. Unlike the traditional framework, in which the principal's goal is to maximize expected payoff (of outputs) minus transfers, I assume the principal's sole objective is to maximize output, constrained by a budget limit. The budget with a specific usage is a common practice in firms. Anthony et al. (2007) and Horngren (2009) provide comprehensive empirical evidence for this approach. In firms, managers (the principals in my model) control the resources without being the ultimate owners, giving rise to agency costs, known as residual loss in finance (Jensen and Meckling, 1976). To combat this loss, a budget constraint with a specific usage is commonly deployed.

Suppose that an agent chooses an effort level $e \geq 0$, which can be discrete or continuous. A continuous output $y \geq 0$ is random and distributed according to a probability density function $g(y; e)$. A principal's objective is to maximize the agent's output, subject to a budget constraint equals to 1. The transfer $t(y)$ depends on the output y, such that $0 \leq t(y) \leq 1$. A cutoff transfer with parameter d is of the form $t(y) = 1$ if $y \geq d$ and $t(y) = 0$ otherwise. The agent's cost of effort is given by a lower semi-continuous function $c(e)$ and he maximizes the expected transfer minus the cost.

Corollary 3. *Suppose that g satisfies the monotone likelihood ratio condition: for*

all $y_1 \leq y_2$, $\frac{g(y_2;e)}{g(y_1;e)}$ *is weakly increasing in* e*. Then, there exists an optimal transfer that is a cutoff transfer.*[22]

This is an immediate corollary of Corollary 1. My model can be "projected down" onto this classic setting. The argument involves making the signal s publicly observable, which removes the truthful report issue. Then set output $y = 1/|s - \theta|$ and effort $e = \lambda$. I formalize a sense that my principal-expert setting is strictly richer than the classic principal-agent setting and derive new results for the classic setting. Economically, this corollary suggests that my insights apply not only to contracting for modern "knowledge/information economy" jobs but also for classic "production/manufacturing" jobs.

The literature on principal-agent problems has long been interested in when simple contracts are optimal (Carroll, 2015; Herweg, Müller, and Weinschenk, 2010; Oyer, 2000). This corollary provides an alternative explanation for simple contracts: the optimality of cutoff transfers stems from the nature of budgets in many contracting scenarios. The bounded budget that can only be used for this task can lead to the optimality of cutoff transfers, and the monotone likelihood ratio property is a sufficient condition to generate this sharp characterization.

Classic principal-agent models maximize the principal's expected payoff (of outputs) minus transfers, subject to incentive compatibility (IC), individual rationality (IR), and sometimes limited liability constraints. Generally, closed-form solutions for the optimal transfers are not obtainable (see Bolton and Dewatripont, 2004, Chapter 4.5). In a special case with binary effort levels and a risk-neutral principal, a well-known result emerges: the monotone likelihood ratio property is necessary and sufficient for optimal transfers to increase with output (see Laffont and Martimort, 2009, Proposition 4.6). However, analyzing the general case with more than two effort levels is much more challenging. Grossman and Hart (1983) study this problem with continuous effort under the assumption of finite output levels. Their findings show that even the monotone likelihood ratio property is not sufficient to ensure

[22] Banks and Sundaram (1998) study how a long-lived principal interacts with a series of short-lived agents with moral hazard and adverse selection. Agents have different types (abilities) and can derive utility from working. Each agent can work for at most two periods. The principal designs a retention rule that maps from the first-period output to two outcomes: retain or fire. The principal aims to maximize output. Banks and Sundaram (1998) show that a cutoff retention rule can be optimal under some condition on the agent's preference and MLRP. However, the principal is unable to induce effort with moral hazard alone. See their proposition 3.5 for a detailed discussion.

optimal transfers to be increasing. Beyond some weak predictions, we do not have a characterization of optimal transfers.[23]

From a methodological perspective, my results rely on techniques of monotone comparative statics, which allows me to impose relatively mild assumptions on the output distribution. Another popular method is the first-order approach (Rogerson, 1985; Jewitt, 1988). To ensure the validity of the first-order approach, people either impose restrictive conditions on the distribution of outputs, like the convexity of output distribution (Rogerson, 1985), or impose strong assumptions on the agent's utility function (Jewitt, 1988). My results do not rely on the first-order approach.

[23]For a risk-neutral principal, we only know that the optimal transfer cannot be decreasing everywhere, nor can it increase faster than the output everywhere.

1.7 The Assumption That Transfer Depends Only on the Difference Between the State and Report

In this section, I show that given the uniform prior, it is without loss to assume that the transfer depends only on the difference $a - \theta$. To make the expected transfer well-defined for the uniform prior, I take two approaches. In the first approach, I define the agent's expected transfer to be conditional on the event that the signal falls in some set of finite Lebegue measure. Then I show that the conclusion holds for any such set. The conditioning on some set is merely an artifact to deal with the improperness of unbounded uniform prior. In the second approach, I study an alternative setting where the state space is a circle in $\mathbb{R}^2$ and the prior is proper and uniform. This unit circle is similar to the one-point compactification of $\mathbb{R}$.

Throughout this section, I allow the transfer to depend on both the state and the report. To ensure that the agent's report is well-defined, I assume that the family of maps $a \mapsto t(y+a, a)$ parametrized by y is equicontinuous. This is trivially satisfied when the transfer depends only on $\theta - a$, as $t(y + a, a)$ stays constant as a varies. When the prior is improper uniform on $\mathbb{R}$, I additionally assume that $t(\theta, \theta + x)$ vanishes uniformly (in θ) as x tends to infinity.

First, I show that the agent's report is well-defined.

Lemma 3. *The maximum*

$$\max_{a \in \mathbb{R}} \int_{\mathbb{R}} t(\theta', a)\varphi(\theta'; s, \lambda)d\theta'$$

is well-defined.

Proof of Lemma 3. Fix a signal s and precision λ. Let $T(a)$ be the agent's expected transfer after reporting a,

$$T(a) = \int t(\theta', a)\varphi(\theta'; s, \lambda)d\theta'.$$

If $T(a) = 0$ for all $a \in \mathbb{R}$, then the maximum is well-defined. Suppose that there exists $a_1 \in \mathbb{R}$ such that $T(a_1) > 0$. As $t(\theta, \theta + x)$ vanishes uniformly (in θ) as x tends to infinity, there exists a $K_1 > 0$ such that for all $|\theta - a| > K_1$, $t(\theta, a) < \frac{T(a_1)}{2}$. Moreover, there exists a $K_2 > 0$ such that $\Phi(\lambda K_2) > 1 - \frac{T(a_1)}{4}$. Let $K = \max(K_1, K_2)$.

If $|a-s|>2K$,

$$\begin{aligned}
T(a) =& \int t(\theta', a)\varphi(\theta'; s, \lambda)d\theta' \\
=& \int_{\theta':|\theta'-s|<K} t(\theta', a)\varphi(\theta'; s, \lambda)d\theta' + \int_{\theta':|\theta'-s|\geq K} t(\theta', a)\varphi(\theta'; s, \lambda)d\theta' \\
\leq& \frac{T(a_1)}{2}\int_{\theta':|\theta'-s|<K} \varphi(\theta'; s, \lambda)d\theta' + 1 \cdot \int_{\theta':|\theta'-s|\geq K} \varphi(\theta'; s, \lambda)d\theta' \\
\leq& \frac{T(a_1)}{2} + 2[1-\Phi(\lambda K)] \\
=& T(a_1).
\end{aligned}$$

Thus,

$$\max_{a\in\mathbb{R}} T(a) = \max_{|a-s|\leq 2K} T(a).$$

Next, I show that $T(\cdot)$ is continuous.

$$\begin{aligned}
T(a+\epsilon) - T(a) =& \int t(\theta', a+\epsilon)\varphi(\theta'; s, \lambda)d\theta' - \int t(\theta', a)\varphi(\theta'; s, \lambda)d\theta' \\
=& \int [t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)]\varphi(\theta'; s, \lambda)d\theta' + \int [t(\theta'-\epsilon, a) - t(\theta', a)]\varphi(\theta'; s, \lambda)d\theta' \\
=& \int [t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)]\varphi(\theta'; s, \lambda)d\theta' + \int t(\theta', a)[\varphi(\theta'+\epsilon; s, \lambda) - \varphi(\theta'; s, \lambda)]d
\end{aligned}$$

$$\begin{aligned}
|T(a+\epsilon) - T(a)| \leq& \sup_{\theta'} |t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)| + \left| \int t(\theta', a)[\varphi(\theta'; s-\epsilon, \lambda) - \varphi(\theta'; s, \lambda)]d\theta' \right| \\
\leq& \sup_{\theta'} |t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)| + \int |\varphi(\theta'; s-\epsilon, \lambda) - \varphi(\theta'; s, \lambda)|d\theta' \\
\leq& \sup_{\theta'} |t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)| + 2\int_{\theta'\leq s-\frac{\epsilon}{2}} \varphi(\theta'; s-\epsilon, \lambda) - \varphi(\theta'; s, \lambda)d\theta' \\
\leq& \sup_{\theta'} |t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)| + 2[\Phi(\frac{\epsilon}{2}\lambda) - \Phi(-\frac{\epsilon}{2}\lambda)]
\end{aligned}$$

Note that $\sup_{\theta'} |t(\theta', a+\epsilon) - t(\theta'-\epsilon, a)|$ tends to 0 as ϵ tends to 0 by the equicontinuity of t. Moreover, the CDF Φ is continuous. Thus, T is continuous. A continuous function achieves a maximum on a compact set. So the maximum is well-defined.

□

Next, I show that it is without loss to assume that the transfer depends only on the difference $\theta - a$. For a transfer rule $t(\theta - a)$ that depends only on the difference $\theta - a$, let $\hat{E}(\lambda; t)$ denote the agent's expected transfer when he has to report truthfully,

$$\hat{E}(\lambda; t) = \int_{\mathbb{R}} t(\theta')\varphi(\theta'; 0, \lambda)d\theta'.$$

Let $\hat{\boldsymbol{\lambda}}(t)$ denote the corresponding induced precision,

$$\hat{\boldsymbol{\lambda}}(t) = \begin{cases} \max[\arg\max \hat{E}(\cdot;t) - c(\cdot)], & \text{if } \max \hat{E}(\cdot;t) - c(\cdot) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For a transfer rule $t(\theta, a)$ that depends on both the state and the report, we need to integrate the signal when computing the expected transfer. Let g denote the PDF of the signal, which is also uniform. Fix any measureable set A of finite Lebegue measure. I define the agent's expected transfer conditional on the signal falls in set A, i.e., $s \in A$

$$E(\lambda;t) = \int_{s\in A} g(s)ds \left(\max_a \int t(\theta', a)\varphi(\theta'; s, \lambda)d\theta' \right)$$

$$\boldsymbol{\lambda}(t) = \begin{cases} \max[\arg\max E(\cdot;t) - c(\cdot)], & \text{if } \max E(\cdot;t) - c(\cdot) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 4. *For a transfer rule $t(\theta, a)$, there exists a transfer rule $t^*(\theta - a)$ that depends only on the difference $\theta - a$; given transfer rule t^*, the agent chooses precision $\boldsymbol{\lambda}(t)$ when he is forced to reveal the signal. That is,*

$$\boldsymbol{\lambda}(t) = \hat{\boldsymbol{\lambda}}(t^*).$$

Proof. Let $\boldsymbol{a}(s;\lambda)$ denote the agent's report after observing signal s given precision λ.

$$\boldsymbol{a}(s;\lambda) = \arg\max_a \int_{\mathbb{R}} t(\theta', a)\varphi(\theta'; s, \lambda)d\theta'.$$

$$E(\lambda;t) = \int_{s\in A} g(s)ds \int t(\theta', \boldsymbol{a}(s;\lambda))\varphi(\theta'; s, \lambda)d\theta'$$

Note that this integral is well-defined since $\boldsymbol{a}(s;\lambda)$ is well-defined and the set A has finite Lebegue measure. Let $\theta'' = \theta' - s$,

$$E(\lambda;t) = \int_{s\in A} g(s)ds \int t(\theta'' + s, \boldsymbol{a}(s;\lambda))\varphi(\theta''; 0, \lambda)d\theta''.$$

Let $t^*(\theta'') = \int_{s\in A} g(s)t(\theta'' + s, \boldsymbol{a}(s;\boldsymbol{\lambda}(t)))ds$. By Fubini-Tonelli Theorem,

$$\begin{aligned} E(\boldsymbol{\lambda}(t);t) &= \int_{s\in A} g(s)ds \int t(\theta'' + s, \boldsymbol{a}(s;\boldsymbol{\lambda}(t)))\varphi(\theta''; 0, \boldsymbol{\lambda}(t))d\theta'' \\ &= \int_{\mathbb{R}} t^*(\theta'')\varphi(\theta''; 0, \boldsymbol{\lambda}(t))d\theta'' \\ &= \hat{E}(\boldsymbol{\lambda}(t); t^*). \end{aligned}$$

Moreover, at $\lambda \neq \boldsymbol{\lambda}(t)$,

$$\begin{aligned} E(\lambda; t) &= \int_{s\in A} g(s)ds \int_{\mathbb{R}} t(\theta', \boldsymbol{a}(s;\lambda))\varphi(\theta'; s, \lambda)d\theta' \\ &\geq \int_{s\in A} g(s)ds \int_{\mathbb{R}} t(\theta', \boldsymbol{a}(s;\boldsymbol{\lambda}(t)))\varphi(\theta'; s, \lambda)d\theta' \\ &= \int_{s\in A} g(s)ds \int t(\theta'' + s, \boldsymbol{a}(s;\boldsymbol{\lambda}(t)))\varphi(\theta''; 0, \lambda)d\theta'' \\ &= \int_{\mathbb{R}} t^*(\theta'')\varphi(\theta''; 0, \lambda)d\theta'' \\ &= \hat{E}(\lambda; t^*) \end{aligned}$$

where the inequality follows by the definition of $\boldsymbol{a}(s;\lambda)$. By the definition of $\boldsymbol{\lambda}(t)$, $\hat{\boldsymbol{\lambda}}(t^*) = \boldsymbol{\lambda}(t)$. □

Given this Lemma, we can input the constructed t^* into the proof of Theorem 1, where I show that there exists a cutoff transfer that induces a higher precision.

In the second approach, the state space is a circle in $\mathbb{R}^2$ with a circumference of 1, and the prior distribution is uniform. This setup corresponds to cases where the state is periodic, such as the unit vector in $\mathbb{R}^2$ which corresponds a direction in $\mathbb{R}^2$ (In Machina triangle, each direction corresponds to an indifference curve.), the time of day (e.g., 15:20) or the day of the year (e.g., October 19). The model remains almost the same. The only exception is that ϕ admits a compact support $[-M, M]$ and the precision $\lambda \in [2M, +\infty]$. The proof that it is without loss to assume that transfer depends only on the difference $\theta - a$ is almost the same as above. The exception is that the integral of s in Lemma 4 can be taken over the entire circle.

1.8 Omitted Proofs

Proof of Lemma 1.

$$\frac{\partial \ln \phi(\lambda x)}{\partial \ln \lambda} = \frac{\phi'(\lambda x)}{\phi(\lambda x)}\lambda x = -\eta(\lambda x).$$

Thus, we have

$$\frac{\partial[\ln \frac{\phi(\lambda x_1)}{\phi(\lambda x_2)}]}{\partial \ln \lambda} = \eta(\lambda x_2) - \eta(\lambda x_1).$$

The second part follows by

$$\frac{\partial[\frac{\varphi(x_1;0,\lambda)}{\varphi(x_2;0,\lambda)}]}{\partial \lambda} = \frac{\partial[\frac{\phi(\lambda x_1)}{\phi(\lambda x_2)}]}{\partial \lambda}.$$

□

Proof of Lemma 2. Let Φ denote the anti-derivative of ϕ, i.e., $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$. We compute the expected transfer given the cutoff transfer d

$$E(\lambda; d) = 2\Phi(\lambda d) - 1.$$

The derivative of the expected transfer with respect to λ is

$$\frac{\partial E(\lambda; d)}{\partial \lambda} = 2d\phi(\lambda d).$$

Let us see how the derivative changes when varying d,

$$\frac{\partial \frac{\partial E(\lambda;d)}{\partial \lambda}}{\partial d} = 2\phi(\lambda d)[1 - \eta(\lambda d)].$$

□

Lemma 5. *The agent's expected transfer*

$$E(\lambda; t) = \max_{a \in \mathbb{R}} \int_{\mathbb{R}} t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta'$$

is well-defined.

Proof of Lemma 5. I show that the maximum is well-defined, given that the transfer rule t vanished at infinity. Fix a signal s and precision λ. Let $T(a)$ be the agent's expected transfer after reporting a,

$$T(a) = \int t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta'.$$

If $T(a) = 0$ for all $a \in \mathbb{R}$, the $\arg\max_a T(a)$ is well-defined. Suppose that there exists $a_1 \in \mathbb{R}$ such that $T(a_1) > 0$. As t vanished at infinity, there exists a $K_1 > 0$ such that for all $|x| > K_1$, $t(x) < \frac{T(a_1)}{2}$. Moreover, there exists a $K_2 > 0$ such that $\Phi(\lambda K_2) > 1 - \frac{T(a_1)}{4}$. Let $K = \max(K_1, K_2)$. If $|a - s| > 2K$,

$$\begin{aligned}
T(a) =& \int t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' \\
=& \int_{\theta':|\theta'-s|<K} t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' + \int_{\theta':|\theta'-s|\geq K} t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' \\
\leq& \frac{T(a_1)}{2} \int_{\theta':|\theta'-s|<K} \varphi(\theta'; s, \lambda)d\theta' + 1 \cdot \int_{\theta':|\theta'-s|\geq K} \varphi(\theta'; s, \lambda)d\theta' \\
\leq& \frac{T(a_1)}{2} + 2[1 - \Phi(\lambda K)] \\
=& T(a_1).
\end{aligned}$$

Thus,

$$\arg\max_{a\in\mathbb{R}} T(a) = \arg\max_{|a-s|\leq 2K} T(a).$$

Next, I show that $T(\cdot)$ is continuous.

$$\begin{aligned}
|T(a+\epsilon) - T(a)| =& \left| \int t(\theta' - a - \epsilon)\varphi(\theta'; s, \lambda)d\theta' - \int t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' \right| \\
=& \left| \int t(\theta' - a)\varphi(\theta' + \epsilon; s, \lambda)d\theta' - \int t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' \right| \\
=& \left| \int t(\theta' - a)\varphi(\theta'; s - \epsilon, \lambda)d\theta' - \int t(\theta' - a)\varphi(\theta'; s, \lambda)d\theta' \right| \\
=& \left| \int t(\theta' - a)[\varphi(\theta'; s - \epsilon, \lambda) - \varphi(\theta'; s, \lambda)]d\theta' \right| \\
\leq& \int \left| \varphi(\theta'; s - \epsilon, \lambda) - \varphi(\theta'; s, \lambda) \right| d\theta' \\
=& 2\int_{\theta'\leq s - \frac{\epsilon}{2}} \varphi(\theta'; s - \epsilon, \lambda) - \varphi(\theta'; s, \lambda)d\theta' \\
=& 2[\Phi(\frac{\epsilon}{2}\lambda) - \Phi(-\frac{\epsilon}{2}\lambda)]
\end{aligned}$$

Thus, T is continuous as the CDF Φ is continuous. A continuous function achieves a maximum on a compact set. So the maximum in $E(\lambda; t)$ is well-defined. □

Let $\hat{E}(\lambda; t)$ denote the agent's expected transfer when he has to report truthfully

$$\hat{E}(\lambda; t) = \int_{\mathbb{R}} t(\theta')\varphi(\theta'; 0, \lambda)d\theta'.$$

Let $\hat{\boldsymbol{\lambda}}(t)$ denote the corresponding induced precision

$$\hat{\boldsymbol{\lambda}}(t) = \begin{cases} \max[\arg\max \hat{E}(\cdot; t) - c(\cdot)], & \text{if } \max \hat{E}(\cdot; t) - c(\cdot) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 6. *Suppose f satisfies increasing elasticity above 1. Suppose the agent has to report truthfully. Given a symmetric transfer rule $\tilde{t}$ and induced $\lambda^* = \hat{\boldsymbol{\lambda}}(\tilde{t})$, the new transfer*

$$\tilde{t}'(x) = \begin{cases} 1, & \text{if } \lambda^*|x| < \eta^{-1}(1), \\ \tilde{t}(x), & \text{otherwise} \end{cases}$$

can induce a weakly higher precision $\hat{\boldsymbol{\lambda}}(\tilde{t}') \geq \hat{\boldsymbol{\lambda}}(\tilde{t})$.

Proof of Lemma 6. Let $\Delta t = \tilde{t}' - \tilde{t}$. I shall show that

$$\hat{E}(\lambda^*; \Delta t) \geq \hat{E}(\lambda; \Delta t) \quad \text{for all } \lambda \leq \lambda^*.$$

This implies that $\hat{\lambda}(\tilde{t}') \geq \hat{\lambda}(\tilde{t})$ since $\hat{\lambda}(t) = \max[\arg\max \hat{E}(\cdot; t) - c(\cdot)]$.

Consider two cutoff transfers $0 < d_1 < d_2 \leq \eta^{-1}(1)/\lambda^*$. For all $\lambda \leq \lambda^*$, we have $\eta(\lambda d_1) \leq 1, \eta(\lambda d_2) \leq 1$. By Lemma 2,

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \geq 0 \quad \Leftrightarrow \quad \eta(\lambda d) \leq 1.$$

This implies $\frac{\partial E(\lambda;d)}{\partial \lambda}$ increases in $d \in [d_1, d_2]$ for all $\lambda \leq \lambda^*$, which implies $E(\lambda; d_2) - E(\lambda; d_1)$ increases in $\lambda \in [0, \lambda^*]$. As Δt is an integral of such difference of cutoff transfers and $\hat{E}(\lambda; t)$ is linear in t, $\hat{E}(\lambda; \Delta t)$ increases in $\lambda \in [0, \lambda^*]$.

□

Lemma 7. *Suppose f satisfies increasing elasticity above 1. Let $\eta^{-1}(1) < x_1 < x_2$. Then*

$$\frac{\phi(\lambda' x_2)}{\phi(\lambda' x_1)} \geq \frac{\phi(x_2)}{\phi(x_1)} \quad \textit{for all} \quad \lambda' < 1.$$

Proof of Lemma 7. By Lemma 1,

$$\ln \frac{\phi(\lambda' x)}{\phi(x)} = \int_{\lambda'}^{1} \eta(\lambda x) d \ln \lambda.$$

Then,

$$\frac{\phi(\lambda' x_1)}{\phi(x_1)} \leq \frac{\phi(\lambda' x_2)}{\phi(x_2)} \quad \text{for all} \quad \lambda' < 1$$

is equivalent to

$$\int_{\lambda'}^{1} \eta(\lambda x_1) d \ln \lambda \leq \int_{\lambda'}^{1} \eta(\lambda x_2) d \ln \lambda \quad \text{for all} \quad \lambda' < 1.$$

Let $y = \ln \lambda$, then it suffices to show

$$\int_{\ln \lambda'}^{0} \eta(x_1 \exp(y)) dy \leq \int_{\ln \lambda'}^{0} \eta(x_2 \exp(y)) dy \quad \text{for all} \quad \lambda' < 1.$$

Let $\Delta y = \ln(x_2/x_1) > 0$. Then it suffices to show

$$\int_{\ln \lambda'}^{0} \eta(x_2 \exp(y - \Delta y)) dy \leq \int_{\ln \lambda'}^{0} \eta(x_2 \exp(y)) dy \quad \text{for all} \quad \lambda' < 1.$$

Let $g(y) = \eta(x_2 \exp(y))$. Then it suffices to show

$$\int_{\ln \lambda'}^{0} g(y - \Delta y) dy \leq \int_{\ln \lambda'}^{0} g(y) dy \quad \text{for all} \quad \lambda' < 1. \tag{1.4}$$

Note that function $g(y)$ is increasing in $y \in [\ln \frac{\eta^{-1}(1)}{x_2}, 0]$ and single-crosses 1 from below at $\ln \frac{\eta^{-1}(1)}{x_2}$.

If $\ln \lambda' \geq -\Delta y$, then for all $y \in [\ln \lambda', 0]$,

$$g(y) \geq g(\ln \lambda') \geq g(-\Delta y) \geq g(y - \Delta y).$$

Suppose $\ln \lambda' < -\Delta y$. Equation (1.4) is

$$\int_{\ln \lambda' - \Delta y}^{-\Delta y} g(y)dy \leq \int_{\ln \lambda'}^{0} g(y)dy \quad \text{for all} \quad \lambda' < 1.$$

I can subtract $\int_{\ln \lambda'}^{-\Delta y} g(y)dy$ from both sides. Then the inequality becomes

$$\int_{\ln \lambda' - \Delta y}^{\ln \lambda'} g(y)dy \leq \int_{-\Delta y}^{0} g(y)dy \quad \text{for all} \quad \lambda' < 1.$$

It holds since for all $y \in [-\Delta y, 0]$, $y' \in [\ln \lambda' - \Delta y, \ln \lambda']$,

$$g(y) \geq g(-\Delta y) \geq g(\ln \lambda') \geq g(y').$$

□

Proof of Theorem 1 (3) implies (1) and Proof of Theorem 2. Suppose that f satisfies increasing elasticity above 1. If there is no t such that $\lambda(t) > 0$, then the problem is trivial.[24] Now suppose there exists a t such that $\lambda(t) > 0$. In step 1, I shall show that for all transfer t with $\lambda(t) > 0$, I can construct a new cutoff transfer d such that $\lambda(d) \geq \lambda(t)$. That is, cutoff transfer d weakly improves over t. In step 2, I shall optimize over cutoff transfers and prove Theorem 2. In step 3, I show uniqueness under stronger conditions.

Step 1: Optimality of Cutoff Transfers

Fix a transfer rule t. First, I show that there exists a transfer rule $t^*(\theta - a)$ such that given t^*, the agent chooses precision $\lambda(t)$ when he has to report truthfully. That is,

$$\hat{\lambda}(t^*) = \lambda(t).$$

For the transfer rule t that depends on the state and the report, Lemma 4 gives us the desired t^*. Now assume that t depends only on the difference between the state and

[24] In this case, we adopt the convention that every transfer is optimal.

the report, $\theta - a$. The construction of t^* is simpler and is as follows. Let $\boldsymbol{a}(s;\boldsymbol{\lambda}(t))$ denote the agent's report after observing signal s at the precision $\boldsymbol{\lambda}(t)$.

$$\boldsymbol{a}(s;\boldsymbol{\lambda}(t)) = \arg\max_a \int_{\mathbb{R}} t(\theta' - a)\varphi(\theta'; s, \boldsymbol{\lambda}(t))d\theta'.$$

As both t and φ are translation invariant, $\boldsymbol{a}(s;\boldsymbol{\lambda}(t)) - s$ is a constant. I define a new transfer rule t^* by

$$t^*(x) = t(x + s - \boldsymbol{a}(s;\boldsymbol{\lambda}(t))).$$

Given transfer t^*, the agent reports truthfully at $\boldsymbol{\lambda}(t)$. Note that replacing t by t^* only changes the agent's report by a constant. Two transfers t and t^* provide the same incentive for the agent to choose λ, i.e.,

$$E(\cdot;t) = E(\cdot;t^*), \quad \boldsymbol{\lambda}(t^*) = \boldsymbol{\lambda}(t).$$

I define an auxiliary problem: the agent has to reveal his signal. The agent's expected transfer is

$$\hat{E}(\lambda;t^*) = \int_{\mathbb{R}} t^*(\theta')\varphi(\theta'; 0, \lambda)d\theta'.$$

It must be true that for all $\lambda \in \mathbb{R}_+$,

$$\hat{E}(\lambda;t^*) \leq E(\lambda;t^*) \tag{1.5}$$

and

$$\hat{E}(\boldsymbol{\lambda}(t);t^*) = E(\boldsymbol{\lambda}(t);t^*). \tag{1.6}$$

The inequality holds as the agent loses the flexibility to misreport in the auxiliary problem. The equality holds as the agent reports truthfully at $\lambda = \boldsymbol{\lambda}(t)$. By Equation (1.5) and (1.6),

$$\hat{\boldsymbol{\lambda}}(t^*) = \boldsymbol{\lambda}(t^*) = \boldsymbol{\lambda}(t).$$

Then, I symmetrify the transfer rule t^* to obtain

$$\tilde{t}(x) = \frac{1}{2}[t^*(x) + t^*(-x)].$$

For all $\lambda \in \mathbb{R}_+$, we have

$$\hat{E}(\lambda;\tilde{t}) = \hat{E}(\lambda;t^*).$$

$$\hat{\boldsymbol{\lambda}}(\tilde{t}) = \hat{\boldsymbol{\lambda}}(t^*) = \boldsymbol{\lambda}(t).$$

By Lemma 6, I can augment transfer $\tilde{t}$ to a new transfer $\tilde{t}'$

$$\tilde{t}'(x) = \begin{cases} 1, & \text{if } \boldsymbol{\lambda}(t)|x| < \eta^{-1}(1), \\ \tilde{t}(x), & \text{otherwise} \end{cases}$$

and induces higher precision

$$\hat{\boldsymbol{\lambda}}(\tilde{t}') \geq \hat{\boldsymbol{\lambda}}(\tilde{t}) = \boldsymbol{\lambda}(t).$$

Then I can construct a cutoff transfer d. I pin down the cutoff by requiring that cutoff transfer d and $\tilde{t}'$ offers the same expected transfer at $\boldsymbol{\lambda}(t)$, i.e.,

$$E(\boldsymbol{\lambda}(t); d) = \hat{E}(\boldsymbol{\lambda}(t); \tilde{t}'). \tag{1.7}$$

Let $\Delta t(x) = 1_{|x|\leq d} - \tilde{t}'(x)$ denote the transfer difference. The transfer difference must take the form

$$\Delta t(x) \begin{cases} = 0, & \text{if } |x| \leq \eta^{-1}(1)/\boldsymbol{\lambda}(t), \\ \geq 0, & \text{if } \eta^{-1}(1)/\boldsymbol{\lambda}(t) < |x| \leq d, \\ \leq 0, & \text{otherwise.} \end{cases}$$

Compare the expected transfer given cutoff transfer d and the expected transfer $\tilde{t}'$ in the auxiliary problem.

$$E(\lambda; d) - \hat{E}(\lambda; \tilde{t}') = \int_{\mathbb{R}} \Delta t(\theta')\varphi(\theta'; 0, \lambda)d\theta'.$$

Consider y_1, y_2 such that $\eta^{-1}(1)/\boldsymbol{\lambda}(t) < y_1 < y_2$. Let $x_1 = \boldsymbol{\lambda}(t)y_1$, $x_2 = \boldsymbol{\lambda}(t)y_2$. Then $\eta^{-1}(1) < x_1 < x_2$. By Lemma 7, for all $\lambda < \boldsymbol{\lambda}(t)$, we have

$$\frac{\phi(x_2\lambda/\boldsymbol{\lambda}(t))}{\phi(x_1\lambda/\boldsymbol{\lambda}(t))} \geq \frac{\phi(x_2)}{\phi(x_1)}.$$

$$\frac{\phi(\lambda y_2)}{\phi(\lambda y_1)} \geq \frac{\phi(\boldsymbol{\lambda}(t)y_2)}{\phi(\boldsymbol{\lambda}(t)y_1)}.$$

$$\frac{\varphi(y_2; 0, \lambda)}{\varphi(y_1; 0, \lambda)} \geq \frac{\varphi(y_2; 0, \boldsymbol{\lambda}(t))}{\varphi(y_1; 0, \boldsymbol{\lambda}(t))}.$$

This implies

$$E(\lambda; d) - \hat{E}(\lambda; \tilde{t}') \leq 0, \text{ if } \lambda < \boldsymbol{\lambda}(t).$$

As $\hat{\boldsymbol{\lambda}}(\tilde{t}') \geq \boldsymbol{\lambda}(t)$ and the definition $\boldsymbol{\lambda}(t) = \max[\arg\max E(\cdot; t) - c(\cdot)]$, we have $\boldsymbol{\lambda}(d) \geq \boldsymbol{\lambda}(t)$ (see Figure 1.8).

Step 2: Proof of Theorem 2

Next, I optimize over cutoff transfers to prove Theorem 2 and show the existence of the optimal cutoff transfer. If we set $d < \bar{d}$, the agent never chooses to work. So consider $d \geq \bar{d}$.

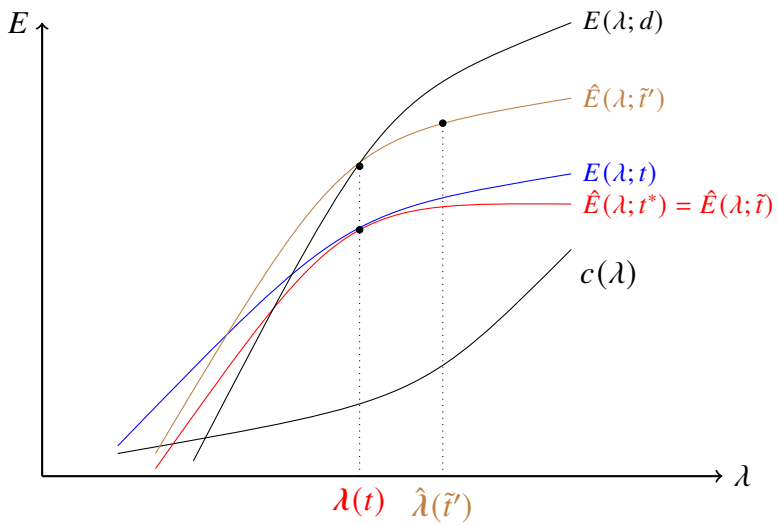


Figure 1.8: Expected Transfer

Case 1: $\lambda(\bar{d})\bar{d} \geq \eta^{-1}(1)$.

Let $\tilde{\lambda} = \eta^{-1}(1)/\bar{d} \leq \lambda(\bar{d})$. For any $d > \bar{d}$, if $\lambda(d) \leq \tilde{\lambda}$, then $\lambda(d) \leq \lambda(\bar{d})$. Consider $\lambda(d) > \tilde{\lambda}$. For all $d \geq \bar{d}$ and $\lambda \geq \tilde{\lambda}$,

$$\eta(\lambda d) \geq \eta(\bar{d}\tilde{\lambda}) \geq 1 \quad \Rightarrow \quad \frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \leq 0.$$

Since $\lambda(d) = \max[\arg\max E(\cdot; d) - c(\cdot)]$. By the Topkis' monotone comparative statics theorem, $\lambda(d) \leq \lambda(\bar{d})$ for all $d \geq \bar{d}$. Thus, $\bar{d}$ is the optimal cutoff.

Case 2: $\lambda(\bar{d})\bar{d} < \eta^{-1}(1)$**.** For all $\bar{d} \leq d \leq d^*$ and $\lambda \leq \frac{\eta^{-1}(1)}{d}$,

$$\eta(\lambda d) \leq 1 \quad \Rightarrow \quad \frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \geq 0.$$

This implies $\lambda(d)$ is increasing in d for $\bar{d} \leq d \leq d^*$. Now suppose $d > d^*$. If $\lambda(d)d \leq \eta^{-1}(1)$, then

$$\lambda(d) \leq \eta^{-1}(1)/d \leq \eta^{-1}(1)/d^* = \lambda^*.$$

If $\lambda(d)d > \eta^{-1}(1)$, for all $d \geq d^*$ and $\lambda \geq \eta^{-1}(1)/d$,

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \leq 0 \quad \Rightarrow \quad \lambda(d) \leq \lambda(d^*)$$

by the Topkis' monotone comparative statics theorem.

Now, I prove existence. In case 1, $\bar{d}$ is the optimal cutoff transfer. In case 2, I can increase d starting from $\bar{d}$. Before hitting d^*, $\lambda(d)$ increases in d. As $\eta^{-1}(1)$ is well-defined and finite given that f satisfies increasing elasticity above 1. Eventually, increasing $\lambda(d)d$ can hit $\eta^{-1}(1)$. Thus, d^* is well defined.

Step 3: Uniqueness

Finally, suppose that ϕ satisfies strictly increasing elasticity above 1, the cost function is continuously differentiable, and the optimal precision is an interior solution (not 0 or $+\infty$).[25] I show that the optimal transfer rule is unique. First, I show that for any transfer rule t that is not a cutoff transfer, the cutoff transfer constructed above induces a strictly larger precision: $\boldsymbol{\lambda}(d) > \boldsymbol{\lambda}(t)$. Since t is not a cutoff transfer, at least one of $\tilde{t}' - \tilde{t}$ and Δt is a non-zero function. If $\tilde{t}' - \tilde{t}$ is non-zero,

$$\left.\frac{\partial[\hat{E}(\lambda;\tilde{t}') - \hat{E}(\lambda;\tilde{t})]}{\partial\lambda}\right|_{\lambda=\lambda^*} > 0. \tag{1.8}$$

This follows by a strict version of Lemma 6 in which strictly increasing elasticity above 1 implies

$$\eta(\lambda d) < 1 \quad \Rightarrow \quad \frac{\partial^2 E(\lambda;d)}{\partial\lambda\partial d} > 0.$$

As $\hat{E}(\cdot;\tilde{t}) - c(\cdot)$ is continuously differentiable,[26] $\hat{\boldsymbol{\lambda}}(\tilde{t})$ is interior, Equation (1.8) implies $\hat{\boldsymbol{\lambda}}(\tilde{t}') > \hat{\boldsymbol{\lambda}}(\tilde{t})$ by Edlin and Shannon, 1998. This implies $\boldsymbol{\lambda}(d) > \boldsymbol{\lambda}(t)$.

If Δt is non-zero, next I show

$$\left.\frac{\partial[E(\lambda;d) - \hat{E}(\lambda;\tilde{t}')]}{\partial\lambda}\right|_{\lambda=\boldsymbol{\lambda}(t)} > 0. \tag{1.9}$$

By the construction of Δt,

$$\int \Delta t(x)\varphi(x;0,\boldsymbol{\lambda}(t))dx = 0 \tag{1.10}$$

$$\int \Delta t(x)\phi(\boldsymbol{\lambda}(t)x)dx = 0.$$

Take some small $\epsilon > 0$. Let $\lambda' = \boldsymbol{\lambda}(t)\exp(\epsilon)$. Consider

$$E(\lambda';d) - \hat{E}(\lambda';\tilde{t}') = \int \Delta t(x)\lambda'\phi(\lambda'x)dx.$$

By $\ln(\lambda') - \ln(\boldsymbol{\lambda}(t)) = \epsilon$ and Lemma 1,

$$\ln\phi(\lambda'x) - \ln\phi(\boldsymbol{\lambda}(t)x) = -\eta(\boldsymbol{\lambda}(t)|x|))\epsilon + o(\epsilon).$$

[25] Distribution ϕ satisfies *strictly increasing elasticity above 1* if $\eta(\cdot)$ single-crosses 1 from below, $\{x|\eta(x) = 1\}$ is a singleton, and is strictly increasing after the cross.

[26] $\hat{E}(\cdot;\tilde{t})$ is continuously differentiable as ϕ is continuously differentiable.

Thus,

$$\begin{aligned} E(\lambda';d) - \hat{E}(\lambda';\tilde{t}') &= \int \Delta t(x)\lambda'\phi(\lambda' x)dx \\ &= \lambda' \int \Delta t(x)\phi(\boldsymbol{\lambda}(t)x)\exp(-\epsilon\eta(\boldsymbol{\lambda}(t)|x|))dx + o(\epsilon) \\ &= -\lambda' \int \Delta t(x)\phi(\boldsymbol{\lambda}(t)x)\epsilon\eta(\boldsymbol{\lambda}(t)|x|)dx + o(\epsilon), \end{aligned}$$

where the last equality follows by Equation (1.10).

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{E(\lambda';d) - \hat{E}(\lambda';\tilde{t}')}{\epsilon} &= -\boldsymbol{\lambda}(t) \int \Delta t(x)\phi(\boldsymbol{\lambda}(t)x)\eta(\boldsymbol{\lambda}(t)|x|)dx \\ &= -\int \Delta t(x)\varphi(x;0,\boldsymbol{\lambda}(t))\eta(\boldsymbol{\lambda}(t)|x|)dx \end{aligned}$$

$$\left.\frac{\partial[E(\lambda;d) - \hat{E}(\lambda;\tilde{t}')]}{\partial\lambda}\right|_{\lambda=\boldsymbol{\lambda}(t)} = -\frac{1}{\boldsymbol{\lambda}(t)} \int \Delta t(x)\varphi(x;0,\boldsymbol{\lambda}(t))\eta(\boldsymbol{\lambda}(t)|x|)dx$$

which is strictly positive by Equation (1.10), $\eta(y)$ being strictly increasing for $y > \eta^{-1}(1)$, and $\Delta t(x)$ being supported on $|x| \geq \frac{\eta^{-1}(1)}{\boldsymbol{\lambda}(t)}$. Again by $\hat{E}(\cdot;\tilde{t}') - c(\cdot)$ being continuously differentiable, $\hat{\boldsymbol{\lambda}}(\tilde{t}')$ is interior, Equation (1.9) implies $\boldsymbol{\lambda}(d) > \hat{\boldsymbol{\lambda}}(\tilde{t}') \geq \boldsymbol{\lambda}(t)$ by Edlin and Shannon (1998).

I have shown that for any transfer rule t that is not a cutoff transfer, $\boldsymbol{\lambda}(d) > \boldsymbol{\lambda}(t)$. Thus, any optimal transfer must be a cutoff transfer. Given strictly increasing elasticity above 1, cost function being continuously differentiable, optimal precision being interior, all comparative statics analyses in the proof of Theorem 2 are strict. Thus, there exists a unique optimal transfer rule, which is a cutoff transfer.

□

Lemma 8. *Suppose that for all cost functions, there exists an optimal transfer that is a cutoff transfer. For a cost function $c \in C^1$, let d^* be the optimal cutoff and λ^* be the induced precision with $\eta(\lambda^* d^*) \geq 1$. Then for all $x_1 \in [0, \lambda^* d^*)$ and $x_2 \in [\lambda^* d^*, +\infty)$,*

$$\eta(x_1) \leq \eta(x_2).$$

Proof of Lemma 8. Without loss I can assume $c(\lambda^*) = E(\lambda^*; d^*)$ since I can increase c by a constant without affecting d^* and λ^*. Similarly, without loss of generality, I assume that λ^* uniquely maximizes $E(\lambda; d^*) - c(\lambda)$.[27]

[27] Otherwise, I can always increase $c(\lambda)$ for $\lambda \neq \lambda^*$, without affecting d^* and λ^*.

Towards a contradiction, suppose $\exists x_1 \in [0, \lambda^* d^*)$, $x_2 \in [\lambda^* d^*, +\infty)$ such that $\eta(x_1) > \eta(x_2)$. As $\eta(\cdot)$ is continuous, I can pick $\delta_1 > 0$, $\delta_2 > 0$ small enough such that for all $y_1 \in [\frac{x_1}{\lambda^*} - \delta_1, \frac{x_1}{\lambda^*} + \delta_1]$ and $y_2 \in [\frac{x_2}{\lambda^*} - \delta_2, \frac{x_2}{\lambda^*} + \delta_2]$,

$$\eta(y_1 \lambda^*) > \eta(y_2 \lambda^*) \tag{1.11}$$

and

$$\int_{\frac{x_1}{\lambda^*} - \delta_1}^{\frac{x_1}{\lambda^*} + \delta_1} \varphi(x; 0, \lambda^*) dx = \int_{\frac{x_2}{\lambda^*} - \delta_2}^{\frac{x_2}{\lambda^*} + \delta_2} \varphi(x; 0, \lambda^*) dx.$$

I define

$$\Delta t(x) = -1_{|x| \in [\frac{x_1}{\lambda^*} - \delta_1, \frac{x_1}{\lambda^*} + \delta_1]} + 1_{|x| \in [\frac{x_2}{\lambda^*} - \delta_2, \frac{x_2}{\lambda^*} + \delta_2]}$$

and a new transfer $\underline{t}(x) = 1_{|x| \leq d^*} + \Delta t(x)$.

I can set δ_1 and δ_2 to be small enough such that the agent reports truthfully given the transfer $\underline{t}$, since $\varphi'(d^*; 0, \lambda^*) < 0$ is bounded from above due to $\eta(\lambda^* d^*) \geq 1$. To see this, it suffices to consider the case $\lambda^* = 1$. If the agent reports truthfully, the expected transfer is $2\Phi(d^*) - 1$. Now suppose the agent misreports by $\epsilon' > 0$. Then his expected transfer is

$$\begin{aligned}
&\Phi(d^* - \epsilon') + \Phi(d^* + \epsilon') - 1 \\
&+\Phi(x_2 - \epsilon' + \delta_2) - \Phi(x_2 - \epsilon' - \delta_2) + \Phi(x_2 + \epsilon' + \delta_2) - \Phi(x_2 + \epsilon' - \delta_2) \\
&-[\Phi(x_1 - \epsilon' + \delta_1) - \Phi(x_1 - \epsilon' - \delta_1)] - [\Phi(x_1 + \epsilon' + \delta_1) - \Phi(x_1 + \epsilon' - \delta_1)],
\end{aligned}$$

where the second and the third line can be made arbitrarily close to 0 as δ_1 and δ_2 tend to 0, since F is continuous. Moreover, the difference

$$\Phi(d^* - \epsilon') + \Phi(d^* + \epsilon') - 2\Phi(d^*)$$

is strictly negative and is strictly decreasing in ϵ', due to f being single-peaked and $\eta(d^*) = -\frac{f'(d^*)}{\phi(d^*)} d^* \geq 1$. Therefore, the expected transfer under misreport is lower than $2\Phi(d^*) - 1$ when δ_1 and δ_2 are small enough.

Then we have

$$E(\lambda^*; \underline{t}) = E(\lambda^*; d^*).$$

By Lemma 1 and Equation (1.11),

$$\left. \frac{\partial [E(\lambda; \underline{t}) - E(\lambda; d^*)]}{\partial \lambda} \right|_{\lambda = \lambda^*} > 0.$$

Since previously λ^* uniquely maximizes $E(\lambda; d^*) - c(\lambda)$, we can pick δ_1 and δ_2 small enough such that $\boldsymbol{\lambda}(\underline{t}) \in (\lambda^* - \epsilon, \lambda^* + \epsilon)$, for some small $\epsilon > 0$. As c is continuously differentiable, $E(\lambda^*; \underline{t}) = E(\lambda^*; d^*)$,

$$\left.\frac{\partial[E(\lambda; \underline{t}) - E(\lambda; d^*)]}{\partial \lambda}\right|_{\lambda=\lambda^*} > 0,$$

by Edlin and Shannon, 1998 we have $\boldsymbol{\lambda}(\underline{t}) > \lambda^*$, contradicting that (d^*, λ^*) is optimal (Figure 1.9).

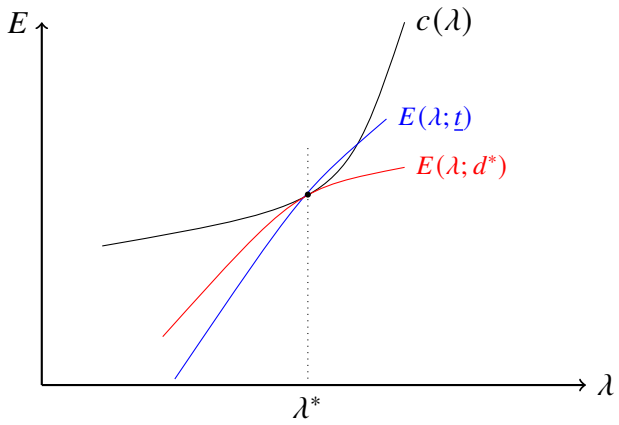


Figure 1.9: Expected Transfer

□

Proof of Theorem 1 (2) implies (3). Suppose that for all increasing and continuously differentiable cost functions, there exists an optimal transfer that is a cutoff transfer. Take a increasing cost function $c_0 \in C^1$, with d_0 and λ_0 being the corresponding optimal cutoff and induced precision. I can pick c_0 such that λ_0 is the unique maximizer of[28] $E(\cdot; d_0) - c_0(\cdot)$. As $E(\cdot; d_0) - c(\cdot)$ is continuously differentiable, I can use the first-order approach. As

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \leq 0 \quad \Leftrightarrow \quad \eta(\lambda d) \geq 1,$$

we have $\eta(\lambda_0 d_0) \geq 1$. If not, I can slightly increase d_0, which leads to a larger

$$\left.\frac{\partial E(\lambda; d)}{\partial \lambda}\right|_{\lambda=\lambda_0}.$$

[28]Given any increasing cost function $c_0 \in C^1$, I can always construct a new increasing and continuously differentiable cost function by increasing $c_0(\lambda)$ for all $\lambda \neq \lambda_0$ while keeping $c_0(\lambda_0)$ and $\frac{dc_0}{d\lambda}|_{\lambda=\lambda_0}$ unchanged.

This induces a larger $\lambda > \lambda_0$, a contradiction.[29]

Since the cutoff transfer d_0 is the optimal transfer and $c_0 \in C^1$, by Lemma 8 and $\eta(\cdot)$ being continuous,

$$\eta(x) \leq \eta(\lambda_0 d_0) \quad \text{if } x \leq \lambda_0 d_0$$

$$\eta(x) \geq \eta(\lambda_0 d_0) \quad \text{if } x \geq \lambda_0 d_0.$$

Note that $\eta(0^+) < 0$ (as f is integrable around 0), $\eta(\lambda_0 d_0) \geq 1$, $\eta(\cdot)$ is continuous, I can find the largest x where $\eta(\cdot)$ crosses 1 from below

$$x_1 = \min\{x | \eta(x') \geq 1, \text{ if } x' \geq x\} \leq \lambda_0 d_0.$$

$\eta(x) \geq 1$ when $x \geq x_1$.

Now pick d_1 and λ_1 such that $\lambda_1 d_1 = x_1$. Construct an increasing cost function $c_1 \in C^1$ such that

$$c_1(\lambda) > E(\lambda; d_1) \quad \text{if } \lambda \neq \lambda_1$$

$$c_1(\lambda_1) = E(\lambda_1; d_1)$$

$$\left.\frac{\partial E(\lambda; d_1)}{\partial \lambda}\right|_{\lambda=\lambda_1} = \left.\frac{dc_1(\lambda)}{d\lambda}\right|_{\lambda=\lambda_1}.$$

As

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \leq 0 \quad \Leftrightarrow \quad \eta(\lambda d) \geq 1$$

and $\eta(x) \geq 1$ for all $x \geq x_1$, the cutoff transfer d_1 is the best among all cutoff transfers for cost function c_1. Since there exists an optimal transfer that is a cutoff transfer, cutoff transfer d_1 is the optimal transfer, and λ_1 is the maximum precision. By Lemma 8,

$$\eta(x) \leq 1 \quad \text{if } x \leq x_1$$

$$\eta(x) \geq 1 \quad \text{if } x \geq x_1.$$

Thus, $\eta(\cdot)$ single-crosses 1 at x_1.

Similarly, for all $x_2 > x_1$, I can pick d_2 and λ_2 such that $\lambda_2 d_2 = x_2$. Pick a increasing cost function $c_2 \in C^1$ such that

$$c_2(\lambda) > E(\lambda; d_2) \quad \text{if } \lambda \neq \lambda_2$$

$$c_2(\lambda_2) = E(\lambda_2; d_2)$$

[29] As λ_0 is the unique maximizer of $E(\cdot; d_0) - c_0(\cdot)$, I can always make the increment on d_0 small enough such that the new maximizer never falls in $[0, \lambda_0)$.

$$\left.\frac{\partial E(\lambda; d_2)}{\partial \lambda}\right|_{\lambda=\lambda_2} = \left.\frac{dc_2(\lambda)}{d\lambda}\right|_{\lambda=\lambda_2}.$$

By the same argument, we have

$$\eta(x) \le \eta(x_2) \quad \text{if } x \le x_2$$

$$\eta(x) \ge \eta(x_2) \quad \text{if } x \ge x_2.$$

Thus, $\eta(\cdot)$ is increasing once it goes above 1 at x_1.

□

Proof of Proposition 1. First suppose for all λ, $c_2(\lambda) \ge E(\lambda; d^*(c_1))$. Then $\bar{d}(c_2) \ge d^*(c_1)$, implying $d^*(c_2) \ge d^*(c_1)$. Moreover, since

$$\text{for all } \lambda \ge \lambda^*(c_1) \text{ and } d \ge d^*(c_1), \quad \frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \le 0$$

and $c_2 - c_1$ is increasing,

$$\max[\arg\max_{\lambda} E(\lambda; d^*(c_2)) - c_2(\lambda)] \le \max[\arg\max_{\lambda} E(\lambda; d^*(c_1)) - c_1(\lambda)].$$

Thus, $\lambda^*(c_2) \le \lambda^*(c_1)$. Now suppose for some λ, $c_2(\lambda) \le E(\lambda; d^*(c_1))$. Note that

$$\begin{aligned}
&\max[\arg\max_{\lambda} E(\lambda; d^*(c_1)) - c_2(\lambda)] \\
=&\max[\arg\max_{\lambda} E(\lambda; d^*(c_1)) - c_1(\lambda) - (c_2(\lambda) - c_1(\lambda))] \\
\le&\max[\arg\max_{\lambda} E(\lambda; d^*(c_1)) - c_1(\lambda)] = \lambda^*(c_1),
\end{aligned}$$

where the inequality follows by $c_2 - c_1$ being increasing. Thus,

$$\boldsymbol{\lambda}(d^*(c_1); c_2) d^*(c_1) \le \lambda^*(c_1) d^*(c_1) = \eta^{-1}(1).$$

By the proof of Theorem 2, $d^*(c_2) \ge d^*(c_1)$ and $\lambda^*(c_2) \le \lambda^*(c_1)$.

□

Proof of Corollary 2. Changing the noise from ε_1 to $k\varepsilon_1$ is equivalent to keeping the noise at ε_1 and changing the cost function from $c_1(\lambda)$ to $c_2(\lambda) = c_1(k\lambda)$. Notice that

$$c_2(\lambda) - c_1(\lambda) = c_1(k\lambda) - c_1(\lambda)$$

which is increasing in λ due to the convexity of c_1. The conclusion follows by Proposition 1. □

Proof of Proposition 2. The proof of the first part remains the same as Theorem 1. For proving the second part, the only difference is to note that the expected transfer is

$$E(\lambda; d) = \int_0^{\lambda d} \phi(r) \frac{\pi^{n/2}}{\Gamma(n/2+1)} dr^n,$$

where Γ is the gamma function and $\frac{\pi^{n/2}}{\Gamma(n/2+1)} r^n$ is the volume of the n-dimensional ball.

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} = \frac{n\pi^{n/2}}{\Gamma(n/2+1)} \phi(\lambda d)(\lambda d)^{n-1}[n - \eta(\lambda d)].$$

Thus,

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} \geq 0 \quad \Leftrightarrow \quad \eta(\lambda d) \leq n.$$

The rest of the argument remains the same.

□

Proof of Proposition 3. The optimality of cutoff transfers follows by the Proof of Theorem 1 (3) implies (1), as Gaussian distributions satisfy increasing elasticity.

For proving the counterpart of Theorem 2, it suffices to notice that

$$\frac{\partial^2 E(\lambda; d)}{\partial \lambda \partial d} = \frac{n\pi^{n/2}}{\Gamma(n/2+1)} \phi(\Lambda(\lambda)d)(\Lambda(\lambda)d)^{n-1} \frac{d\Lambda(\lambda)}{d\lambda}[n - \eta(\Lambda(\lambda)d)],$$

where $\Lambda(\lambda) = (\frac{1}{\lambda_p^2} + \frac{1}{\lambda^2})^{-\frac{1}{2}}$ for uniform prior and $\Lambda(\lambda) = (\frac{1}{\lambda_p^2} + \frac{1}{\lambda^2+\lambda_0^2})^{-\frac{1}{2}}$ for Gaussian prior. □

Proof of Proposition 4. I first prove the case of uniform prior. Given the agent's signal s, his posterior belief about the state is $\mathcal{N}(s, 1/\lambda^2)$. Since the principal's signal is Gaussian centered at θ with variance $\frac{1}{\lambda_p^2}$, the agent's posterior belief about s_p is

$$\mathcal{N}(s, \frac{1}{\lambda^2} + \frac{1}{\lambda_p^2}).$$

Now the agent's payoff maximization problem is same as in Section 1.4 except that s_p replaces θ and his precision is $(\frac{1}{\lambda^2} + \frac{1}{\lambda_p^2})^{-\frac{1}{2}}$. Thus, Proposition 4 follows from Theorem 1 and 2.

The proof for the Gaussian prior is similar. The only difference is that conditional on a signal s, the agent's posterior about s_p is

$$\mathcal{N}(\frac{s\lambda^2}{\lambda_0^2 + \lambda^2}, \frac{1}{\lambda_0^2 + \lambda^2} + \frac{1}{s_p^2}).$$

Thus, the precision is $\Lambda = [\frac{1}{\lambda_0^2+\lambda^2} + \frac{1}{s_p^2}]^{-\frac{1}{2}}$. □

BIBLIOGRAPHY

Anthony, Robert Newton et al. (2007). *Management control systems*. Vol. 12. McGraw-Hill Boston.

Argenziano, Rossella, Sergei Severinov, and Francesco Squintani (2016). "Strategic information acquisition and transmission". In: *American Economic Journal: Microeconomics* 8.3, pp. 119–155.

Banks, Jeffrey S. and Rangarajan K. Sundaram (1998). "Optimal Retention in Agency Problems". In: *Journal of Economic Theory* 82.2, pp. 293–323. ISSN: 0022-0531. DOI: https://doi.org/10.1006/jeth.1998.2422. URL: https://www.sciencedirect.com/science/article/pii/S002205319892422X.

Bolton, Patrick and Mathias Dewatripont (2004). *Contract theory*. MIT press.

Carroll, Gabriel (2015). "Robustness and linear contracts". In: *American Economic Review* 105.2, pp. 536–563.

– (2019). "Robust incentives for information acquisition". In: *Journal of Economic Theory* 181, pp. 382–420.

Chade, Hector and Natalia Kovrijnykh (2016). "Delegated information acquisition with moral hazard". In: *Journal of Economic Theory* 162, pp. 55–92.

Chen, Siyu et al. (2023). "Learning to Incentivize Information Acquisition: Proper Scoring Rules Meet Principal-Agent Model". In: *arXiv preprint arXiv:2303.08613*.

Clark, Aubrey and Giovanni Reggiani (2021). "Contracts for acquiring information". In: *arXiv preprint arXiv:2103.03911*.

Cuccia, Annemarie (2023). *DC left millions unspent from federal grants to end homelessness*. https://thedcline.org/2023/06/08/dc-left-millions-unspent-from-federal-grants-to-end-homelessness/.

Dasgupta, Sulagna (2023). "Optimal Test Design for Knowledge-based Screening". In: *Available at SSRN 4403119*.

Deb, Rahul, Mallesh M Pai, and Maher Said (2018). "Evaluating strategic forecasters". In: *American Economic Review* 108.10, pp. 3057–3103.

– (2023). *Indirect Persuasion*. Centre for Economic Policy Research.

Edlin, Aaron S and Chris Shannon (1998). "Strict monotonicity in comparative statics". In: *Journal of Economic Theory* 81.1, pp. 201–219.

Goel, Prem K and Morris H DeGroot (1979). "Comparison of experiments and information measures". In: *The Annals of Statistics* 7.5, pp. 1066–1077.

Gottlieb, Daniel and Humberto Moreira (2022). "Simple contracts with adverse selection and moral hazard". In: *Theoretical Economics* 17.3, pp. 1357–1401.

Grossman, Sanford J. and Oliver D. Hart (1983). "An Analysis of the Principal-Agent Problem". In: *Econometrica* 51.1, pp. 7–45. (Visited on 10/03/2024).

Hébert, Benjamin and Weijie Zhong (2022). "Engagement maximization". In: *arXiv preprint arXiv:2207.00685*.

Herweg, Fabian, Daniel Müller, and Philipp Weinschenk (Dec. 2010). "Binary Payment Schemes: Moral Hazard and Loss Aversion". In: *American Economic Review* 100.5, pp. 2451–77.

Horngren, Charles T (2009). *Cost accounting: a managerial emphasis*. Pearson Education India.

Jensen, Michael C. and William H. Meckling (1976). "Theory of the firm: Managerial behavior, agency costs and ownership structure". In: *Journal of Financial Economics* 3.4, pp. 305–360. ISSN: 0304-405X. DOI: https://doi.org/10.1016/0304-405X(76)90026-X. URL: https://www.sciencedirect.com/science/article/pii/0304405X7690026X.

Jewitt, Ian (1988). "Justifying the first-order approach to principal-agent problems". In: *Econometrica: Journal of the Econometric Society*, pp. 1177–1190.

Karlin, Samuel and Herman Rubin (1956). "The theory of decision procedures for distributions with monotone likelihood ratio". In: *The Annals of Mathematical Statistics*, pp. 272–299.

Kreutzkamp, Sophie (2023). *Endogenous information acquisition in cheap-talk games*. Tech. rep. Working paper.

Laffont, Jean-Jacques and David Martimort (2009). "The theory of incentives: the principal-agent model". In: *The theory of incentives*. Princeton university press.

Lehmann, Erich Leo (2011). "Comparing location experiments". In: *Selected Works of EL Lehmann*. Springer, pp. 779–791.

Li, Yingkai, Jason D Hartline, et al. (2022). "Optimization of scoring rules". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 988–989.

Li, Yingkai and Jonathan Libgober (2023). "OptimalScoringforDynamicInformation Acquisition". In: *Working Paper*.

Mansouri, Kavahn (2024). *Millions remain unspent in federal funds for homeless students — and time is running out*. https://www.lpm.org/news/2024-07-16/millions-remain-unspent-in-federal-funds-for-homeless-students-and-time-is-running-out.

Milgrom, Paul R (1981). "Good news and bad news: Representation theorems and applications". In: *The Bell Journal of Economics*, pp. 380–391.

National Archives (2024). *Code of Federal Regulations: 886.20 What happens to unused funds from my grant?* https://www.ecfr.gov/current/title-30/chapter-VII/subchapter-R/part-886/section-886.20.

Neyman, Eric, Georgy Noarov, and S Matthew Weinberg (2021). "Binary scoring rules that incentivize precision". In: *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 718–733.

Osband, Kent (1989). "Optimal forecasting incentives". In: *Journal of Political Economy* 97.5, pp. 1091–1112.

Oyer, Paul (2000). "A theory of sales quotas with limited liability and rent sharing". In: *Journal of labor Economics* 18.3, pp. 405–426.

Papireddygari, Maneesha and Bo Waggoner (2022). "Contracts with Information Acquisition, via Scoring Rules". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 703–704.

Rappoport, Daniel and Valentin Somma (2017). "Incentivizing information design". In: *Available at SSRN 3001416*.

Rogerson, William P (1985). "The first-order approach to principal-agent problems". In: *Econometrica: Journal of the Econometric Society*, pp. 1357–1367.

Schick, Allen (2008). *The federal budget: Politics, policy, process*. Brookings Institution Press.

Sharma, Salil, Elias Tsakas, and Mark Voorneveld (2023). *Procuring unverifiable information*. Tech. rep. Mimeo.

Vandeput, Nicolas (2021). *Data science for supply chain forecasting*. de Gruyter.

Whitmeyer, Mark and Kun Zhang (2023). "Buying opinions". In: *arXiv preprint arXiv:2202.05249*.

Wong, Yu Fu (2023). "Dynamic Monitoring Design". In: *Available at SSRN 4466562*.

Zermeno, Luis (2011). "A Principal-Expert Model and the Value of Menus". In: *Working Paper*.

Chapter 2

ESTIMATING NONSEPARABLE SELECTION MODELS: A FUNCTIONAL CONTRACTION APPROACH

2.1 Introduction

Sample selection issues arise when the data available for analysis is not representative of the entire population due to a selection process that systematically excludes certain observations. For example, in consumer demand studies, researchers often only have access to the transaction prices of chosen products, while the prices of non-selected products remain unobserved (Goldberg, 1996; Cicala, 2015; Crawford, Pavanini, and Schivardi, 2018; Allen, Clark, and Houde, 2019; Salz, 2022; Sagl, 2023; Cosconati et al., 2024). Similarly, in auctions, data may include only the winning bids (or certain order statistics), excluding all other submitted bids (Athey and Haile, 2002; Komarova, 2013; Guerre and Y. Luo, 2019; Allen, Clark, Hickman, et al., 2024). Sample selection issues have long been recognized in labor market studies as well. For instance, wage data is typically available only for individuals who choose to work (Gronau, 1974; Heckman, 1974), and, in the original Roy model (Roy, 1951), which examines the occupational distribution of earnings, we observe earnings within an occupation only for those who self-select into working in that sector.

Observing only a selected sample of outcomes—such as prices, bids, or wages—presents significant challenges for estimating two key elements: (1) the model that governs the selection process, such as a consumer demand model, an auction's winning rule, or a labor force participation model; and (2) the distribution of outcomes *prior to selection*, often referred to as "potential outcomes" in the literature. Typically, it is assumed that potential outcomes are generated by an *outcome equation*, which depends on both observable characteristics and unobservable error terms. Flexibly estimating potential outcome distributions is crucial in many empirical contexts, such as analyzing price distributions to understand firms' pricing strategies and wage distributions to examine inequality.

The first solution to sample selection bias is to use maximum likelihood estimation, as in Heckman (1974) and L.-F. Lee (1982, 1983), which relies heavily on distributional assumptions regarding the error terms. More commonly employed methods

for sample selection models are two-step estimators proposed by Heckman (1976, 1979), which introduce a correction term to account for the non-random nature of the sample. A substantial body of theoretical work has been developed to relax the distributional assumptions in the two stages of the estimation procedure (Ahn and Powell, 1993; Andrews and Schafgans, 1998; Chen and Khan, 2003; Das, Newey, and Vella, 2003; Newey, 2007; Newey, 2009; Chernozhukov, Iván Fernández-Val, and S. Luo, 2023). See also Vella (1998) for a comprehensive survey on semi-parametric two-step estimation for selection models.

Our paper proposes a fundamentally different and novel approach to estimating selection models where the outcome equation is nonparametric and nonseparable in error terms. Rather than constructing a reduced-form bias correction term and controlling it in the outcome equation, we directly analyze how the selection model maps the potential outcome distributions to the distributions of selected outcomes and seek to *invert* the mapping. The key insight of our approach is that, given the selection model and potential outcome distributions across all alternatives, we can derive the likelihood of an outcome being selected. Conversely, if this selection likelihood *were* known, we could recover the potential outcome distributions from the observed outcome distributions. This two-way relationship characterizes a fixed-point problem. Building on this intuition, we construct an operator whose fixed point represents the potential outcome distributions and show that this operator is a functional contraction.

Formally, we consider a discrete choice problem in which each alternative is associated with a potential outcome distribution. A selection function maps a vector of realized potential outcomes to a probability distribution over the alternatives. For example, in the consumer demand setting, each alternative represents a product, and the potential outcome is the offered price, with the selection function micro-founded by the consumer's utility maximization problem. We allow the outcome equations to be fully nonparametric with nonseparable error terms and to vary flexibly across different alternatives. We assume that potential outcomes across different alternatives are conditionally independent given observables.

Given the selection function, we construct an operator whose fixed point is the potential outcome distributions. We establish sufficient conditions for it to be a functional contraction (Theorems 3 and 4). Proving contraction within a function space is challenging; to address this, we construct a metric in the same spirit as that in Thompson, 1963. Our results imply that, given the selection function

and the observed distributions of selected outcomes, we can nonparametrically recover the potential outcome distributions. Moreover, this identification result is *constructive*: starting with any initial guess for the potential outcome distributions, we iteratively apply the operator. As the number of iterations approaches infinity, this process converges to the potential outcome distributions associated with the selection function.

We propose a two-step semi-parametric maximum likelihood estimator for the selection function, parameterized by a finite-dimensional parameter, and potential outcome distributions. In the first step, we obtain a nonparametric estimate of the selected outcome distribution directly from the data. Given this estimate, we use our contraction result to recover the potential outcome distributions for *any parameter* in the selection function. In the second step, we construct the model-implied choice probabilities and match them with the data moments. Once we have an estimator for the selection parameter, a plug-in estimator for the potential outcome distribution can be readily obtained.

We establish the consistency and asymptotic normality of the proposed estimator (Theorems 5 and 6). This is particularly challenging because the mapping from the potential to the selected outcome distributions does not have a closed form. We prove that this mapping is a homeomorphism, a key result in establishing consistency and asymptotic normality.

To examine the finite sample properties of our estimator, we conduct Monte Carlo simulations across various designs of the outcome equation. Our results show that the biases in our estimator are generally small, and the standard deviation decreases as the sample size increases across all simulation designs. Our nonparametric estimation of the potential outcome distributions outperforms the standard two-step method when the two-step method misspecifies the outcome equations. Notably, even when the selection function is misspecified by econometricians, our method performs robustly in estimating the potential outcome distributions.

Compared to the traditional two-step method, our approach offers several key advantages. First, we allow for fully nonparametric estimation of potential outcome distributions. Importantly, our approach accommodates nonseparable error terms in the outcome equation, allowing for fully heterogeneous effects of covariates on outcomes. Moreover, we impose no symmetry assumptions, allowing the potential outcome distributions to vary flexibly across alternatives. Unlike most selection correction approaches that focus on estimating conditional mean models (e.g., Das,

Newey, and Vella, 2003 and various other semi-parametric versions)[1], our goal is to recover the *entire outcome distribution* with a flexible specification. We correct for sample selection bias across the entire distribution of potential outcomes by examining how the bias is *systematically* generated by the selection model. More recently, Arellano and Bonhomme (2017) propose a method to correct for sample selection in quantile regression models; see also Newey (2007) and Ivan Fernández-Val, van Vuuren, and Vella (2024) for recent developments in nonseparable sample selection models.

Second, our approach does not require an instrument to exogenously shift the choice probability, a typical requirement in the two-step method to avoid multicollinearity, nor does our approach rely on identification-at-infinity arguments. In practice, finding a suitable instrument can be quite challenging (see Vella, 1998 for further discussion). d'Haultfoeuille and Maurel (2013a) and D'Haultfœuille, Maurel, and Zhang (2018) develop estimation methods for semiparametric sample selection models without an instrument or a large-support regressor, leveraging the independence-at-infinity assumption.

Our approach relies on an alternative assumption: conditional independence of potential outcomes given observables. This assumption is commonly invoked in auction models (e.g., independent private value auctions or mineral rights models)[2] and becomes more plausible when econometricians have access to a rich set of observables. In a binary selection model (e.g., the decision to work) where the potential outcome for one alternative is constant (e.g., the wage for not working is 0) or in censored regression models with a single observed dependent variable, our conditional independence assumption is trivially satisfied. We provide further discussion of this assumption in Section 2.2.

Finally, our method accommodates a flexible selection function, applicable to a variety of empirical settings, including consumer demand, multi-attribute auctions, and labor market decisions. The agent's utility in our model can depend on potential outcomes, observable characteristics, unobserved alternative-specific heterogeneity (such as product quality, compensating differentials, and other nonpecuniary factors), and random preference shocks. Incorporating nonpecuniary components into

[1]These models restrict covariates to affecting only the location of the outcome distribution. A recent paper by Chernozhukov, Iván Fernández-Val, and S. Luo (2023) proposes a semi-parametric generalization of the Heckman selection model which accommodates rich patterns of heterogeneity in the effects of covariates on outcomes and selection.

[2]See Athey and Haile (2007) for further discussion on the conditional independence assumption in auction models and potential testing approaches.

the selection model has proven essential in empirical studies (e.g., Heckman and Sedlacek, 1985; S. T. Berry, 1994; S. Berry, Levinsohn, and Pakes, 1995) and has gained attention in recent theoretical research (Bayer, Khan, and Timmins, 2011; d'Haultfoeuille and Maurel, 2013b; Mourifie, Henry, and Meango, 2020; Canay, Mogstad, and Mountjoy, 2024; J. H. Lee and Park, 2023).

Our method is applicable to a wide range of empirical applications. For example, in a companion paper (Cosconati et al., 2024), we estimate consumer demand in the auto insurance market when only the transaction prices of selected insurance plans are observed. In this market, insurance companies employ risk-based pricing, leading to significant price variation across consumers. Our method enables nonparametric, firm-specific estimates of the offered price distribution, offering valuable insights into the heterogeneity of firms' pricing strategies and, ultimately, the precision of their risk-rating technology. In Section 2.6, we provide a more detailed discussion on applications to three empirical settings: consumer demand, auction models with incomplete bid information, and Roy models in labor economics, along with related literature.

The rest of the paper is organized as follows. Section 2.2 formally introduces our model, with an illustrative example provided at the end. Section 2.3 presents the main theoretical results. In Section 2.4, we describe the semi-parametric maximum likelihood estimator and its asymptotic properties. Section 2.5 reports the results of our Monte Carlo simulations, and Section 2.6 discusses various empirical applications. Finally, Section 2.7 concludes.

2.2 Model

In Sections 2.2–2.3, all analyses are conditional on observable characteristics x, which we omit to simplify notation. Throughout the paper, we use the consumer demand example to illustrate the main results and clarify ideas; however, the approach is broadly applicable to other selection models.

Consider a discrete choice problem. There is a finite set of alternatives $\mathcal{J} = \{1, \cdots, J\}$. Each alternative is associated with a price distribution. Let $G_j \in \Delta([\underline{p}_j, \overline{p}_j])$ represent the price distribution associated with alternative j, where $\Delta(Y)$ denotes the set of all cumulative distribution functions over a set $Y \subset \mathbb{R}$. We assume that $p_j \sim G_j$ are independently distributed across alternatives (conditional on x). The collection of G_j is denoted by $G = \prod_{j \in \mathcal{J}} G_j$. We refer to G as the *offered* price distribution.

A *selection function* is denoted by $f = (f_1, f_2, \cdots, f_J)$ where f_j maps the prices of alternatives $\boldsymbol{p} = (p_1, \cdots, p_J)$ to a strictly positive probability of selecting alternative $j \in \mathcal{J}$.[3] We assume that the selection function is continuously differentiable,

$$f_j \in \mathscr{C}^1 \colon \prod_j [\underline{p}_j, \overline{p}_j] \to (0, 1),$$

with $\sum_{j \in \mathcal{J}} f_j \leq 1$. Here, the inequality allows for the case with an outside option. The selection function is a primitive of the model. To provide a microfoundation, for example, f might be derived from a consumer's utility maximization problem as illustrated in Section 2.2.

Let $\boldsymbol{p}_{-j} = (p_1, \cdots, p_{j-1}, p_{j+1}, \cdots, p_J)$ denote the vector of prices excluding j's price. The probability of selecting j conditional on p_j is given by

$$Pr_j(p_j; G) = \int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k \neq j} dG_k(p_k), \tag{2.1}$$

where $Pr_j(\cdot; G)$ is a function defined on $[\underline{p}_j, \overline{p}_j]$. The assumption that prices are independent across different alternatives allows us to express the joint distribution of $\boldsymbol{p}_{-j}$ as the product of their individual marginal distribution functions.

Let $\tilde{G}_j \in \Delta([\underline{p}_j, \overline{p}_j])$ represent the price distribution conditional on selecting alternative j. We derive $\tilde{G}_j$ using Bayes' rule:

$$\tilde{G}_j(p) = \frac{\int_{\underline{p}_j}^{p} Pr_j(y; G) dG_j(y)}{\int_{\underline{p}_j}^{\overline{p}_j} Pr_j(y; G) dG_j(y)}. \tag{2.2}$$

Note that G_j and $\tilde{G}_j$ share the same support, as selection function f_j is strictly positive. Let $\tilde{G} = \prod_{j \in \mathcal{J}} \tilde{G}_j$ and we call $\tilde{G}$ *selected* price distribution. Equations (2.1) and (2.2) define a mapping from G to $\tilde{G}$. Let $F \colon \prod_j \Delta([\underline{p}_j, \overline{p}_j]) \to \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ denote this mapping, i.e., $\tilde{G} = F(G)$.

In many empirical settings, researchers have access only to the selected price distribution. However, the key primitives of interest are often the offered price distribution. Our research question is how to recover the offered price distribution G from

[3] The assumption that the probability of selecting each alternative is strictly positive is analogous to the overlap assumption in the treatment effect literature, which requires each individual to have a positive probability of receiving each treatment level. This assumption is crucial for recovering the offered price distribution. To illustrate, consider a scenario where $f_j = 0$ whenever p_j falls within a certain subset of $[\underline{p}_j, \overline{p}_j]$. In this case, any p_j within that subset would not be observed in the data, making it impossible to identify G_j within that subset without introducing additional assumptions.

the observed selected price distribution $\tilde{G}$. Note that both G and $\tilde{G}$ are collections of J cumulative distribution functions. Therefore, the cardinality of unknowns and constraints are exactly the same in Equation (2.2) (assuming the selection function is known). Since a cumulative distribution function is an infinite-dimensional object, the key challenge is solving for a collection of infinite-dimensional objects entangled in a nonlinear system. We will explore this in detail in Section 2.3.

An Illustrative Example

We now present a simple example to illustrate the key assumptions of our model and compare them to the standard assumptions in the literature. Consider a consumer choosing between two products, $j = 1, 2$, to maximize her utility. The consumer's utility from product j is given by a scaler value:

$$u_j = \gamma p_j + \varepsilon_j, \tag{2.3}$$

where p_j represents the price of product j for this consumer, and ε_j represents an unobserved utility shock. We abstract from the possibility that the consumer's utility may depend on observable characteristics and unobserved product heterogeneity for this example. In this model, the price sensitivity parameter γ and the distribution of ε_j determine the selection function f. Let $\tilde{\varepsilon} = \varepsilon_1 - \varepsilon_2$ denote the error difference. If $\tilde{\varepsilon} \sim \mathcal{N}(0, 1)$, this represents a binary probit model, and the selection function for product 1 takes the following form:

$$f_1(p_1, p_2) = 1 - \Phi_{\mathcal{N}}\left(\gamma(p_2 - p_1)\right),$$

where $\Phi_{\mathcal{N}}$ denotes the CDF for standard normal distribution.

In this illustrative example, we consider a simple linear outcome equation with an additive error term. For each product $j = 1, 2$, the price is generated by the following equation:

$$p_j = x\beta_j + \eta_j, \tag{2.4}$$

where x represents observable characteristics, and η_j denotes a random shock, which, for simplicity, is assumed to be independent of x.

Suppose the econometrician observes the price of product 1 only when it is chosen

by the consumer. We derive the conditional mean of p_1 given that it is observed:

$$\begin{aligned} E(p_1|x, u_1 > u_2) &= x\beta_1 + E(\eta_1|\gamma p_1 + \varepsilon_1 - (\gamma p_2 + \varepsilon_2) > 0) \\ &= x\beta_1 + E(\eta_1|x \underbrace{\gamma(\beta_1 - \beta_2)}_{\beta^*} + \underbrace{[\gamma(\eta_1 - \eta_2) + \tilde{\varepsilon}]}_{\text{composite error: } \varepsilon^*} > 0) \\ &= x\beta_1 + E(\eta_1|x\beta^* + \varepsilon^* > 0). \end{aligned} \tag{2.5}$$

The conditioning term $x\beta^* + \varepsilon^* > 0$ in Equation (2.5) represents the reduced-form selection model typically seen in the literature. Sample selection issue arises when η_1 and ε^* are correlated, so that $E(\eta_1|x\beta^* + \varepsilon^* > 0) \neq 0$. In the two-step estimation literature, researchers often impose assumptions on the joint distribution of $(\varepsilon^*, \eta_1, \eta_2)$. For example,

$$\begin{bmatrix} \varepsilon^* \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \right).$$

We now take a closer look at the correlation between the composite error (ε^*) and the error in the outcome equation (η_1). Specifically,

$$\begin{aligned} cov(\varepsilon^*, \eta_1) &= cov(\gamma(\eta_1 - \eta_2) + \tilde{\varepsilon}, \eta_1) \\ &= \gamma var(\eta_1) - \gamma cov(\eta_1, \eta_2) + cov(\eta_1, \tilde{\varepsilon}). \end{aligned} \tag{2.6}$$

Equation (2.6) shows that the error term η_1 directly enters the composite error ε^*, implying that $cov(\varepsilon^*, \eta_1) \neq 0$ unless $\gamma = 0$. This correlation is *by construction* in selection models, as agents make decisions after observing the potential outcomes. Another common concern regarding selection bias arises from potential correlation between errors in the outcome equation (e.g., η_1) and those in the structural selection model (e.g., $\tilde{\varepsilon}$), as represented by the third term in Equation (2.6). For example, unobserved productivity factors may create correlation between a worker's willingness to work and their wage. Our model also accommodates this type of correlation.

The only assumption we impose is that the error terms in outcome equations across different alternatives are independent conditional on observables. This implies that $cov(\eta_1, \eta_2) = 0$ in Equation (2.6). In a simple binary model with only one dependent variable of interest, such as Tobit Type 1 or Type 2, this assumption holds trivially. Heckman and Honore (1990) show that, under a strong log-normality assumption, the correlation structure between two outcome variables can be identified; however,

this result does not hold more generally (see discussions in French and Taber, 2011). Due to the nature of the selection problem, the data include only the price of the selected alternative, while competing prices for unselected alternatives are not observed. If the prices of the two products tend to move together, we would not be able to observe this pattern. French and Taber (2011) point out that since the data provides only two one-dimensional price distributions, it is impossible to recover the full joint distribution of a two-dimensional object without imposing additional assumptions.

The conditional independence assumption is commonly employed in auction models, such as independent private value auctions or mineral rights models, where signals are assumed to be independent given the common value. This assumption is more plausible when econometricians have access to a rich set of observables. The conditional independence assumption essentially rules out the presence of a common unobserved factor, x^*, that introduces correlation between outcomes, even after conditioning on observables. When this assumption is not satisfied, the observed price distribution for each alternative, conditional on observable x, is a mixture of price distributions conditional on (x, x^*). We then need to first analyze this mixture model and use additional parametric structures or instruments to identify the selected price distributions conditional on (x, x^*). Techniques for this type of deconvolution problem have been studied in the literature (see the recent survey articles by Compiani and Kitamura, 2016; Hu, 2017) and are beyond the scope of this paper. We maintain the conditional independence assumption for the remainder of the paper.

Finally, we highlight several additional features that differentiate our model from the existing literature. First, our model allows the outcome equation to be fully flexible and nonparametrically specified as $p_j = h_j(x, \eta_j)$, where h_j is an unknown function that may be nonseparable in the error term. Our goal is to recover the entire distribution of p_j conditional on x, rather than only estimating the parameters in the conditional mean function, such as β_j in Equation (2.4). Importantly, we fully account for heterogeneity in the effects of covariates on outcomes. Second, our model does not require an instrument that exogenously shifts choices between alternatives and is excluded from the outcome equation—a critical requirement for identification and estimation in the two-step method. In other words, we allow the same set of observables to enter both the outcome and selection equations. Moreover, we impose minimal assumptions on the selection function. It can accommodate non-

parametric, nonseparable relationships between observable and unobserved errors, offering much greater flexibility than the utility specification in Equation (2.3); in fact, it *does not* even need to be derived from a utility maximization problem. Our framework also allows for alternative-specific unobserved heterogeneity, which is a desirable feature in many empirical contexts.

2.3 Main Results

We define an operator $T\colon \prod_j \Delta([\underline{p}_j, \overline{p}_j]) \to \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ by inverting Equation (2.2). Let $\Psi = (\Psi_1, \Psi_2, \cdots, \Psi_J) \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$.

$$(T\Psi)_j(p) = \frac{\int_{\underline{p}_j}^{p} d\tilde{G}_j(y)/Pr_j(y;\Psi)}{\int_{\underline{p}_j}^{\overline{p}_j} d\tilde{G}_j(y)/Pr_j(y;\Psi)}. \tag{2.7}$$

Note that in Equation (2.7), $Pr_j(y;\Psi)$ is the probability of selecting j conditional on $p_j = y$ and Ψ_{-j}, where Ψ_{-j} is the collection of offered price distributions of all alternatives except j.

An intuitive way to understand the operator in Equation (2.7) is as follows. Suppose we begin with a conjecture for the offered price distribution, denoted by Ψ. Based on this conjecture, we can calculate the probability of selecting alternative j given $p_j = y$, i.e., $Pr_j(y;\Psi)$. Using this selection probability, we can *invert* the observed distribution of selected prices to infer the distribution of offered prices by dividing $d\tilde{G}_j(y)$ by $Pr_j(y;\Psi)$. The denominator in Equation (2.7) serves as a normalizing factor. This process updates the initial conjecture Ψ. If the conjecture Ψ is correct and matches the true distribution G, the update will also equal G. Thus, the offered price distribution G a fixed point of the operator T.

The operator T is a contraction if there exists some real number $0 \leq \rho < 1$ such that for all $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$D(T\Psi, T\Phi) \leq \rho D(\Psi, \Phi),$$

given some metric D.[4] In the reminder of this section, we first construct the metric D and then characterize the modulus ρ. We discuss several special cases of our model at the end.

[4]We adopt the convention that $+\infty$ and $+\infty$ are not comparable, but $c < +\infty$ for any $c \in \mathbb{R}_+$.

Constructing the Metric

We begin by defining a metric in the set of all cumulative distribution functions for an alternative j. Let Ψ_j and Φ_j denote two probability measures in $\Delta([\underline{p}_j, \overline{p}_j])$. Recall that two probability measures Ψ_j and Φ_j are equivalent, denoted $\Psi_j \sim \Phi_j$, if they are absolutely continuous with respect to each other. When $\Psi_j \sim \Phi_j$, the Radon-Nikodym derivative,

$$\frac{d\Psi_j}{d\Phi_j} \colon [\underline{p}_j, \overline{p}_j] \to (0, \infty),$$

exists, as guaranteed by the Radon-Nikodym Theorem. If both Ψ_j and Φ_j have continuous densities, the Radon-Nikodym derivative simplifies to the ratio of densities:

$$\frac{d\Psi_j}{d\Phi_j}(p) = \frac{\Psi_j'(p)}{\Phi_j'(p)}.$$

Note that

$$\Psi_j = \Phi_j \quad \Leftrightarrow \quad \frac{d\Psi_j}{d\Phi_j}(p) = 1 \qquad \Phi_j\text{-a.e.}$$

In the space $\Delta([\underline{p}_j, \overline{p}_j])$, we define a metric $d \colon \Delta([\underline{p}_j, \overline{p}_j]) \times \Delta([\underline{p}_j, \overline{p}_j]) \to [0, +\infty]$ to simplify the analysis.[5]

$$d(\Psi_j, \Phi_j) = \begin{cases} \ln \operatorname{ess\,sup}_{y \in [\underline{p}_j, \overline{p}_j]} \frac{d\Psi_j}{d\Phi_j}(y) + \ln \operatorname{ess\,sup}_{y \in [\underline{p}_j, \overline{p}_j]} \frac{d\Phi_j}{d\Psi_j}(y), & \text{if } \Psi_j \sim \Phi_j, \\ +\infty & \text{otherwise.} \end{cases}$$

Given our operator T in Equation (2.7), for all $\Psi_j, \Phi_j \in \Delta([\underline{p}_j, \overline{p}_j])$,

$$(T\Psi)_j \sim \tilde{G}_j \sim (T\Phi)_j.$$

Thus,

$$d((T\Psi)_j, (T\Phi)_j) = \ln \operatorname*{ess\,sup}_{p_j} \frac{d(T\Psi)_j}{d(T\Phi)_j}(p_j) + \ln \operatorname*{ess\,sup}_{p_j} \frac{d(T\Phi)_j}{d(T\Psi)_j}(p_j).$$

The observed selected price distribution $\tilde{G}_j$ appears in both $(T\Psi)_j$ and $(T\Phi)_j$. As a result, $\tilde{G}_j$ cancels out in the distance above. Moreover, the denominator in our

[5]This metric is a variant of the Thompson metric (Thompson, 1963). The Thompson metric between two functions $s, q \in \mathbb{R}^Y$ is

$$d_{Thompson}(s, q) = \max\{\ln \sup \frac{s(y)}{q(y)}, \ln \sup \frac{q(y)}{s(y)}\}.$$

operator is a normalizing factor, which is also canceled out after we take the sum of log ratios. Consequently, the distance between $(T\Psi)_j$ and $(T\Phi)_j$ only relies on the ratio between selection probabilities:

$$d((T\Psi)_j, (T\Phi)_j) \leq \sup_{p_j} \ln \frac{Pr_j(p_j; \Psi)}{Pr_j(p_j; \Phi)} + \sup_{p_j} \ln \frac{Pr_j(p_j; \Phi)}{Pr_j(p_j; \Psi)},$$

where equality holds when $\tilde{G}_j$ admits full support on $[\underline{p}_j, \overline{p}_j]$.

Next, we define a metric in the space $\prod_j \Delta([\underline{p}_j, \overline{p}_j])$ by taking the maximum distance among all alternatives:

$$D(\Psi, \Phi) = \max_{j \in \mathcal{J}} d(\Psi_j, \Phi_j)$$

for any $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. From now on, we work with the metric space $(\prod_j \Delta([\underline{p}_j, \overline{p}_j]), D)$.

Functional Contraction

For $j \in \mathcal{J}$, we define the *maximum semi-elasticity difference* as

$$M_j = \sup_{p_j, \boldsymbol{p}_{-j}, \boldsymbol{p}'_{-j}} \left| \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} - \frac{\partial \ln f_j(p_j, \boldsymbol{p}'_{-j})}{\partial p_j} \right|. \tag{2.8}$$

The quantity $\frac{\partial \ln f_j}{\partial p_j}$ measures how sensitive the log of the choice probability changes with respective to the price, and therefore represents the semi-elasticity. Let

$$\rho = \frac{J-1}{4} \max_{j \in \mathcal{J}} (\overline{p}_j - \underline{p}_j) M_j.$$

Theorem 3. *If $\rho < 1$, the operator T is a contraction with modulus less than ρ.*

By the Banach fixed point theorem, whenever $\rho < 1$, any selected distribution $\tilde{G}$ corresponds to a unique offered distribution G. Theorem 3 implies that we can nonparametrically identify the potential outcome distributions G from the observed selected outcome distribution $\tilde{G}$, given the selection function f. Moreover, this result provides a constructive method for solving G. Take any $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, by Theorem 3,

$$D(T^n\Psi, G) = D(T^n\Psi, TG) \leq \rho D(T^{n-1}\Psi, G) \leq \rho^{n-1} D(T\Psi, G),$$

where $D(T\Psi, G)$ is finite. This implies

$$\lim_{n \to \infty} D(T^n\Psi, G) = 0.$$

$$\lim_{n\to\infty} T^n\Psi = G.$$

Thus, we can simply take an initial guess for the potential outcome distributions and iteratively apply the operator. As the number of iterations approaches infinity, this process converges to the potential outcome distributions associated with the selection function.

Note that the condition of Theorem 3 is a joint constraint on the selection function and the price range. The bound on the modulus, ρ, consists of the product between the number of alternatives, the price range $\overline{p}_j - \underline{p}_j$, and the maximum semi-elasticity difference.[6] Our condition requires this product to be small. If we expand the support $[\underline{p}_j, \overline{p}_j]$ to $[\underline{p}'_j, \overline{p}'_j]$ where

$$\underline{p}'_j < \underline{p}_j < \overline{p}_j < \overline{p}'_j$$

with $\tilde{G}$ unchanged, ρ becomes weakly larger, which implies now it is more difficult for the operator T to contract. This comparison is intuitive. Since the domain $\prod_{k\neq j} \Delta([\underline{p}_k, \overline{p}_k]) \times \Delta([\underline{p}'_j, \overline{p}'_j])$ is larger, we are considering more collections of probability measures, making it more challenging to control $\frac{D(T\Psi,T\Phi)}{D(\Psi,\Phi)}$ for all Ψ and Φ in this domain.

To understand the maximum semi-elasticity difference M_j in the modulus ρ, consider an extreme case where the choice probabilities do not vary with prices at all, indicating perfectly inelastic demand. In this scenario, there is effectively no selection and the offered price distribution coincides with the selected price distribution. The modulus equals 0 and we obtain the fixed point immediately.

It may be a concern that a large number of alternatives J would result in a large modulus. However, we show that a large number of alternatives could lead to a small maximum semi-elasticity difference. For example, consider the multinomial logit model, arguably the most popular model for discrete choices due to its analytical form and ease of estimation:

$$f_j(p_1, \cdots, p_J) = \frac{\exp(\gamma p_j)}{\sum_{k=1}^{J} \exp(\gamma p_k)},$$

where γ represents the consumer's price sensitivity. We derive the semi-elasticity for the logit model,

$$\frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} = \gamma(1 - f_j(\boldsymbol{p})).$$

[6]Note that by definition ρ is unitless. Changing the unit of price does not affect ρ.

When J is large, the choice probability for each alternative tends to be small, so that the log derivative is approximately equal to γ. As a result, the maximum semi-elasticity difference is close to 0.

The crux and the bulk of the proof for Theorem 3 is to provide a bound on the ratio

$$\sup_{\Psi,\Phi\in\prod_j \Delta([\underline{p}_j,\overline{p}_j])} \frac{D(T\Psi,T\Phi)}{D(\Psi,\Phi)}.$$

This is difficult as the domain of the supreme, $\prod_j \Delta([\underline{p}_j, \overline{p}_j])$, is a large space. For instance, if $J = 10$, the supreme is over 20 functions. In the proof of this theorem in Section 2.9, we employ a technique called a change of measure, also know as the tilted measure, and combine it with insights from transportation problem. The bound ρ is relatively tight: there exist selection functions for which the supremum is arbitrarily close ρ. In Section 2.8, we connect our contraction result with quantal response equilibria (McKelvey and Palfrey, 1995).

Special Cases

Thus far, we have not imposed any structure on the selection function. For a general selection function, we have to take the supreme over $\boldsymbol{p}_{-j}, \boldsymbol{p}'_{-j}$ to compute the maximum semi-elasticity difference. Now we impose an assumption on the selection function to determine where the supreme is attained.

Assumption 1 (Log Supermodularity)**.** For all $j \in \mathcal{J}$ and $p_j \in [\underline{p}_j, \overline{p}_j]$, $\frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j}$ is weakly increasing in each p_k with $k \neq j$.

Given log supermodularity, the maximum semi-elasticity difference is attained at the boundary,

$$M_j = \sup_{p_j} \left| \frac{\partial \ln f_j(p_j, \overline{\boldsymbol{p}}_{-j})}{\partial p_j} - \frac{\partial \ln f_j(p_j, \underline{\boldsymbol{p}}_{-j})}{\partial p_j} \right|.$$

What is left in the definition of maximum semi-elasticity difference is the supreme over p_j. It turns out that we can use $\overline{p}_j - \underline{p}_j$ in the definition of ρ to eliminate the supreme over p_j and give a tighter bound. The result is as follows,

$$\rho^* = \frac{J-1}{4} \max_{j\in\mathcal{J}} [\ln f_j(\overline{\boldsymbol{p}}) - \ln f_j(\underline{p}_j, \overline{\boldsymbol{p}}_{-j}) - \ln f_j(\overline{p}_j, \underline{\boldsymbol{p}}_{-j}) + \ln f_j(\underline{\boldsymbol{p}})].$$

Theorem 4. *Suppose that Assumption 1 holds. If $\rho^* < 1$, the operator T is a contraction with modulus less than ρ^*.*

Under Assumption 1, the modulus ρ^* takes a much simpler form and is straightforward to compute. The log-supermodularity assumption holds in models widely adopted by empirical researchers. For example, the multinomial logit model satisfies Assumption 1. Another example is the binary probit model we describe in Section 2.2. The log-supermodularity condition in Assumption 1 holds for the binary probit model and Theorem 4 applies.[7] However, Assumption 1 may not hold for probit models with three or more alternatives; in such cases, the more general results in Theorem 3 can be applied.

To summarize, our contraction results provide a novel method for identifying the potential outcome distribution from the observed selected outcome distribution, given any selection function f—whether parametric or nonparametric, and regardless of whether it is microfounded in a utility maximization problem. Moreover, the identification is constructive: starting with an initial guess, iterative application of the operator converges to the potential outcome distributions associated with the selection function. These theoretical results are essential for estimating the selection function and potential outcome distributions, which will be discussed in the next section.

2.4 Estimation

Building on the theoretical results in Section 2.3, we now turn to the estimation of the model's primitives. We begin by discussing the estimation of the offered price distribution G when the selection function f is known, followed by the more complex case where both f and G must be jointly estimated.

In the data, for each individual i, we observe their choice, characteristics, and the price of the selected product. Let $y_{ij} = 1$ if j is chosen by i, and 0 otherwise. Since the alternatives are exclusive, $\sum_{j=1}^{J} y_{ij} = 1$. Let x_{ij} represent a vector of observable characteristics. We define $y_i = (y_{i1}, \cdots, y_{iJ})'$ and $x_i = (x_{i1}', \cdots, x_{iJ}')' \in X$. The observed selected prices in the data enable us to estimate $\tilde{G}$ using standard

[7] To see this, we compute the log derivative for the binary probit model:

$$\frac{\partial \ln f_1(p_1, p_2)}{\partial p_1} = \frac{\gamma \phi_{\mathcal{N}}(\Delta)}{1 - \Phi_{\mathcal{N}}(\Delta)},$$

$$\frac{\partial^2 \ln f_1(p_1, p_2)}{\partial p_1 \partial p_2} = \gamma^2 \frac{d}{d\Delta}\left[\frac{\phi_{\mathcal{N}}(\Delta)}{1 - \Phi_{\mathcal{N}}(\Delta)}\right],$$

where $\Delta = \gamma(p_2 - p_1)$ and the term in the square bracket is known as the hazard rate or inverse Mills ratio. As Gaussian satisfies increasing hazard rate (Baricz, 2008), the log-supermodularity condition in Assumption 1 holds.

nonparametric methods. Let $\hat{G}$ denote the estimate of $\tilde{G}$, and $\hat{G}(x)$ denote the estimate conditional on observable x.

Estimation with a Known Selection Function

In Section 2.3, we show that for a given selection function f, the offered price distribution G can be uniquely determined from the selected price distribution $\tilde{G}$, as the number of iterations of the operator T defined in Equation (2.7) goes to infinity. In practice, however, econometricians typically do not observe the true selected price distribution $\tilde{G}$, but rather an estimate $\hat{G}$, which is subject to sampling errors. Moreover, when iterating the operator T to obtain the offered price distribution G, the process stops after a finite number of iterations m. Therefore, our estimation of G contains these two sources of error.

Let $T^m_{\hat{G}}\Psi$ denote our estimator for G, using the estimated selected price distribution $\hat{G}$ and initiating the operator iteration with $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. The distance between our estimator and the true G is bounded by the sum of sampling errors from a finite sample size and the approximation errors from finite iterations, as shown in the following triangular inequality.

$$D(G, T^m_{\hat{G}}\Psi) \leq \underbrace{D(G, F^{-1}(\hat{G}))}_{\text{finite sample size}} + \underbrace{D(F^{-1}(\hat{G}), T^m_{\hat{G}}\Psi)}_{\text{finite iteration}},$$

where F^{-1} denotes the inverse of F. Recall that F is the mapping from G to $\tilde{G}$ defined in Equations (2.1) and (2.2). The inverse mapping, F^{-1}, maps $\tilde{G}$ back to G. Theorem 3 guarantees that we can obtain G from $\tilde{G}$ by iterating the operator T an infinite number of times.

We first focus on the sampling error $D(G, F^{-1}(\hat{G}))$. The next proposition shows that this error goes to zero as $\hat{G}$ converges to $\tilde{G}$.

Proposition 5. *Suppose $\rho < 1$. The mapping F is a homeomorphism. Moreover, both F and F^{-1} are Lipschitz continuous, with Lipschitz constants $1 + \rho$ and $\frac{1}{1-\rho}$, respectively.*

Since F is a homeomorphism, the inverse F^{-1} is well-defined and $G = F^{-1}(\tilde{G})$. Since F^{-1} is continuous, we have

$$F^{-1}(\hat{G}) \xrightarrow{p} F^{-1}(\tilde{G}) \quad \text{as} \quad \hat{G} \xrightarrow{p} \tilde{G}.$$

Moreover, as F^{-1} is Lipschitz continuous, $F^{-1}(\hat{G})$ converges to G at the same rate as $\hat{G}$ converges to $\tilde{G}$.

We now analyze the approximation error $D(T^m_{\hat{G}}\Psi, F^{-1}(\hat{G}))$ due to the finite number of iterations. Note that this error term tends to 0 at speed ρ^m. Thus, if ρ^m decays faster than the convergence rate of $\hat{G}$ to $\tilde{G}$, then $T^m_{\hat{G}}\Psi$ converges to G at the same rate as $\hat{G}$ converges to $\tilde{G}$. We let $m(n)$ express the dependence of the number of iterations on the sample size. The following result summarizes the discussion above.

Corollary 4. *Suppose that $\hat{G} \xrightarrow{p} \tilde{G}$ at a polynomial rate of n^k with $k > 0$. If*

$$\liminf_{n\to+\infty} \frac{m(n)}{\ln n} > k(\ln(1/\rho))^{-1},$$

then $T^{m(n)}_{\hat{G}}\Psi \xrightarrow{p} G$ at rate n^k.

For instance, if the support of $\tilde{G}$ is finite, $\hat{G} \to \tilde{G}$ at rate $\sqrt{n}$. If

$$\lim_{n\to\infty} \rho^{m(n)}\sqrt{n} = 0 \quad \text{or} \quad \liminf_{n\to+\infty} \frac{m(n)}{\ln n} > \frac{1}{2}(\ln(1/\rho))^{-1},$$

$T^m_{\hat{G}}\Psi$ converges to G at rate $\sqrt{n}$.

Estimation with an Unknown Selection Function

We now consider the case where the selection function f is unknown to econometricians and we jointly estimate f and G. As discussed in Section 2.3, given any selection function f, whether parametric or nonparametric, our contraction results provide a straightforward method for recovering the potential outcome distribution from the observed selected outcome distribution. This step utilizes all the information contained in the selected outcome distribution. To further identify and estimate the selection function f, we must leverage additional data, specifically the "market share" of each alternative.

The dimensionality of market shares determines how flexibly we can estimate f. For example, if market shares are observed conditional on continuously distributed covariates, it is possible to estimate a semiparametric single-index model (Ichimura, 1993; Klein and Spady, 1993) for the selection function f. While allowing for a semiparametric or nonparametric selection function is theoretically possible, implementing it would be highly complex and data-intensive. In most empirical settings, market shares are observed conditional on discrete values of covariates. We therefore focus on the case where the selection function is parametrically specified in the estimation.[8]

[8]In our Monte Carlo simulations, we consider a scenario where the selection function is misspecified by the econometrician. We find that our estimates of the potential outcome distributions remain quite robust even when the selection function is misspecified.

We assume that the selection function f is derived from a standard multinomial choice model with an indirect utility given by

$$u_{ij} = v_j(p_{ij}, x_{ij}, \varepsilon_{ij}; \theta),$$

where v_j is a known function parametrized by a finite-dimensional parameter θ; p_{ij} denotes the offered price of alternative j for individual i; the vector of unobserved error terms $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iJ})$ is jointly distributed according to a known distribution. Note that our framework fully allows that the unobserved error term enters the utility function in a nonseparable way. The individual chooses an alternative to maximize utility, and the selection function f is captured by the parameter θ. Let θ_0 denote the true parameter. For example, one commonly used specification is as follows:

$$u_{ij} = \gamma p_{ij} + x_{ij}'\beta + \xi_j + \varepsilon_{ij}, \quad j = 1, 2, \cdots, J,$$

where ξ_j represents a scalar-valued unobserved characteristic of alternative j. In this example, $\theta = (\gamma, \beta, \boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_J)$.

We estimate the parameter θ in the selection function by matching the model-implied choice probabilities to those observed in the data. Specifically, for an individual with observable characteristic x_i, the probability of choosing alternative j is given by the following equation:

$$Prob_j(x; \theta, \hat{G}, m) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta) d\big(T^m_{\hat{G}(x),\theta}\Psi\big)(\boldsymbol{p}), \tag{2.9}$$

where $T^m_{\hat{G}(x),\theta}\Psi$ represents the estimated offered price distribution after iterating the operator T for m steps, starting with the initial value Ψ. The operator is constructed using the estimated selected price distribution conditional on x, denoted by $\hat{G}(x)$, and the selection function parameterized by θ. Note that θ affects the choice probabilities both directly through the selection function and indirectly through the estimated offered price distribution.

Let $z_i = \{x_i, y_i\}$. Given an i.i.d. sample of $\{z_i\}_{i=1}^n$ and a first-step nonparametric estimator $\hat{G}(x)$, we propose a semiparametric maximum likelihood estimator for θ:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{Q}_n(\theta), \tag{2.10}$$

where

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln Prob_j(x_i; \theta, \hat{G}, m(n)). \tag{2.11}$$

Once $\hat{\theta}$ is obtained, a plug-in estimator for G is given by $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$.

Consistency and Asymptotic Normality

In this section, we show that the estimator defined in Equation (2.10) is consistent and asymptotically normal. We maintain the previous assumptions on the selection function: $f_j \in \mathscr{C}^1$: $\prod_j [\underline{p}_j, \overline{p}_j] \to (0,1)$. The additional technical conditions required for the consistency of $\hat{\theta}$ are as follows.

Assumption 2. (i) The space Θ of parameter θ is compact; (ii) for each $x \in X$, the selection function $f(\boldsymbol{p}; x, \theta)$ is jointly continuous in θ and $\boldsymbol{p}$; (iii) the condition in Theorem 3 holds for all $\theta \in \Theta$, that is, $\sup_{\theta\in\Theta} \rho(\theta) \leq \bar{\rho} < 1$ for some $\bar{\rho}$; (iv) the number of iterations $m(n) \to \infty$; (v) $\hat{G} \xrightarrow{p} \tilde{G}$.

Assumption 3 (Identification)**.** There does no exist $\theta' \in \Theta$, $\theta' \neq \theta_0$, offered price distributions G, $G' \in \left(\prod_j \Delta([\underline{p}_j, \overline{p}_j])\right)^X$ such that for all $j \in \mathcal{J}$ and $x \in X$

$$F(G(x); \theta_0) = F(G'(x); \theta'),$$

$$\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta') dG'(x)(\boldsymbol{p}).$$

Assumption 2 (i) and (ii) are standard regularity conditions. Assumption 2 (iii) ensures that for all $\theta \in \Theta$, the operator T is a contraction. Assumption 2 (iv) requires that the number of iterations m tends to infinity, but it does not impose any restrictions on the rate at which m approaches infinity. Assumption 2 (v) ensures that our first-step estimator $\hat{G}$ is consistent. Assumption 3 imposes the identification condition, which requires that there does not exist another parameter that can yield the same selected price distribution and choice probabilities.

Theorem 5 (Consistency)**.** *Under Assumptions 2 and 3,* $\hat{\theta} \xrightarrow{p} \theta_0$, $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi \xrightarrow{p} G$.

Proving this theorem turns out to be challenging. We cannot rely on the standard consistency arguments for maximum likelihood estimators, as $\hat{Q}_n(\theta)$ is not a sample average. Since all data points are already used to estimate $\hat{G}$, each term in $\hat{Q}_n(\theta)$ depends on the entire dataset. Moreover, the number of iterations depends on the sample size n.

To prove consistency, we invoke the fundamental consistency theorem for extremum estimators (Theorem 2.1 in Newey and McFadden, 1994). We construct the true population objective function as follows:

$$Q_0(\theta) = \mathbb{E}_x \sum_{j=1}^{J} \left(\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p}) \right) \ln\left(Prob^*_j(x; \theta, \tilde{G})\right),$$

where $\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p})$ represents the true probability of selecting alternative j conditional on x; and

$$Prob_j^*(x; \theta, \tilde{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta) dF^{-1}(\tilde{G}(x), \theta)(\boldsymbol{p}). \tag{2.12}$$

Equation (2.12) represents the model-implied choice probability for alternative j conditional on x, given the model parameter θ, the true selected price distribution $\tilde{G}$, and as the number of iterations goes to infinity. By the identification condition in Assumption 3, Q_0 is uniquely maximized at θ_0.

Similarly to Section 2.4, there are two sources of error in the sample objective function $\hat{Q}_n(\theta)$ when approximating the true population objective function $Q_0(\theta)$: (1) sampling error, and (2) errors resulting from the finite number of iterations of the operator T. To focus on the sampling error, we construct the following intermediate objective function where the number of iterations m in Equation (2.11) goes to infinity:

$$\hat{Q}_n^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \left(Prob_j^*(x_i; \theta, \hat{G}) \right).$$

We use the homeomorphism in Proposition 5 to show that $\hat{Q}_n^*$ converges pointwise to Q_0 in probability. We then prove that $\hat{Q}_n^*$ is equicontinuous, which ensures its uniform convergence to Q_0. Lastly, we show that $\hat{Q}_n$ converges uniformly in probability to $\hat{Q}_n^*$ as the number of iterations approaches infinity, which implies that $\hat{Q}_n$ converges uniformly in probability to Q_0, a key to establishing the consistency result. Further details of each step can be found in Section 2.9.

Next, we show that the estimator defined in Equation (2.10) is asymptotically normal. Motivated by our discussion above, we first study the behavior of the estimator when m tends to infinity for each n. Let

$$\hat{\theta}^* = \arg\max_{\theta} \hat{Q}_n^*(\theta),$$

$$\mathfrak{g}^*(z_i; \theta, \hat{G}) = \nabla_\theta \left(\sum_{j=1}^{J} y_{ij} \ln Prob_j^*(x_i; \theta, \hat{G}) \right),$$

where ∇_θ denote the gradient operator with respect to θ. The estimator $\hat{\theta}^*$ solves the first-order condition

$$\frac{1}{n} \sum_{i=1}^{n} \mathfrak{g}^*(z_i; \hat{\theta}^*, \hat{G}) = 0.$$

Proving the asymptotic normality of a semiparametric two-step estimator typically requires a first-order expansion around the nonparametric estimator (see Theorem 8.1 in Newey and McFadden, 1994). In our case, this involves expanding the equation above around $\hat{G}$. A standard argument would apply if $\hat{G}$ entered directly into Equation (2.12). However, it enters through F^{-1}, for which we lack an analytic form. As a result, continuing to work with an infinite-dimensional distribution $\tilde{G}$ becomes extremely challenging.

To make the analysis tractable, we assume that the support of $\tilde{G}$ is finite. This assumption is practically innocuous, as nonparametric estimators are always represented as finite-dimensional vectors in numerical applications. For instance, in consumer demand estimation, $\tilde{G}$ represents a distribution over prices, which are measured in discrete units (e.g., cents), so this assumption is reasonable.

Assumption 4. (i) $\text{supp}(\tilde{G})$ is finite. (ii) θ_0 is in the interior of Θ. (iii) f is twice continuously differentiable in θ. (iv) $\mathbb{E}\nabla_\theta \mathfrak{g}^*(z;\theta_0,\tilde{G})$ is nonsingular. (v) The number of iterations satisfies $\liminf_{n\to+\infty} \frac{m(n)}{\ln n} > \frac{1}{2}(\ln(1/\bar{\rho}))^{-1}$.

Assumption 4(ii)–(iv) are standard regularity conditions. Assumption 2–4(iv) ensure that the estimator $\hat{\theta}^*$ is asymptotically normal. Assumption 4(v) requires that the number of iterations increases rapidly enough for the error introduced by finite iterations to become negligible compared to the error of $\hat{\theta}^*$. Particularly, it guarantees $\sqrt{n}(\hat{\theta} - \hat{\theta}^*) \xrightarrow{p} 0$, which gives us the next result.

Theorem 6 (Asymptotic Normality)**.** *Suppose that Assumption 2, 3, and 4 hold. Then $\hat{\theta}$ is asymptotically normal and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$.*[9] *$T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$ converges to G in probability at rate $\sqrt{n}$.*

2.5 Monte Carlo Simulations

To examine how our estimator for θ and the offered price distribution may perform in practice, we conduct a Monte Carlo simulation experiment with $J = 2$. The utility individual i derives from the two alterantives are specified as follows:

$$u_{i1} = -\gamma \log(p_{i1}) + \xi_1 + \beta x_{i1} + \varepsilon_i,$$
$$u_{i2} = -\gamma \log(p_{i2}) + \xi_2,$$

where p_{ij} and ξ_j are, respectively, the offered price and unobserved heterogeneity for alternative j; $x_{i1} \in \{0, 1\}$ is a binary observable with $Pr(x_{i1} = 1) = 0.5$ that shifts

[9] See the analytical form of V in the proof of Lemma 15.

individual i's choice probabilities; and $\varepsilon_i \sim N(0,1)$ is the error term. Throughout the simulation exercises, we set the utility parameters as follows: $\gamma = 1$, $\xi_1 = 0$, $\xi_2 = 1, \beta = 0.5$. Let $y_i \in \{1,2\}$ denote the choice of individual i.

We consider five data generating processes for the offered prices. Let x_{i2} denote the observable characteristic of individual i that enters the pricing equation. For simplicity, we also restrict x_{i2} to take binary values from $\{0,1\}$, with $Pr(x_{i2} = 1) = 0.7$.

DGP 1: $\log(p_{ij}) = \delta_{0j} + \delta_j x_{i2} + \eta_{ij}$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim N(0, 0.2)$.

DGP 2: $\log(p_{ij}) = \delta_{0j} + \delta_j x_{i2} + \eta_{ij}$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim EV(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim EV(0, 0.2)$.

DGP 3: $\log(p_{ij}) = (\delta_{0j}+\delta_j x_{i2})(1+\eta_{ij})$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim N(0, 0.3)$.

DGP 4: $\log(p_{ij}) = \exp\big((\delta_{0j} + \delta_j x_{i2})(1 + \eta_{ij})\big)$, where $\delta_{01} = 0.2, \delta_1 = 0.1, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 0.3, \eta_{i2} \sim N(0, 0.2)$.

DGP 5: $\log(p_{ij}) = (\delta_{0j} + \delta_j x_{i2})(1 + \eta_{ij})^{-1}$, where $\delta_{01} = 0.2, \delta_1 = 0.1, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 0.3, \eta_{i2} \sim N(0, 0.2)$.

In DGP 1, the error term in the pricing equation is additively separable and follows a normal distribution, which is commonly assumed in empirical applications. DGP 2 assumes instead that the error term follows an extreme value distribution, while in DGP 3, we relax the homoskedasticity assumption. Finally, DGPs 4 and 5 consider scenarios where the pricing function takes a nonseparable form.[10]

For each DGP, we simulate offered prices and individual choices, and assume that the econometricians observe $(y_i, x_{i1}, x_{i2}, p_i)$, where p_i is the price of the chosen alternative. We then apply our method from Section 2.4 to estimate the parameters of the selection function, i.e., $\theta = (\gamma, \xi_2, \beta)$ with ξ_1 normalized to 0, along with the offered price distribution for each alternative.[11] For comparison, we employ the

[10] Although all the offer price distributions admit unbounded support, in simulation we shall assume that the realized price range coincides with the true price range. Given a large sample size, the realized price range supports almost all the probability mass of the offered price distribution. Later we show that the estimation of the offered price distribution performs well.

[11] We estimate the cumulative distribution function of prices at 300 grid points.

classic two-step method, assuming that the pricing equations are linearly separable, with an error term that is independent of x_{i2} and normally distributed. Under this assumption, the two-step method misspecifies the pricing equation under DGPs 2–5. For each design, we run 500 simulations of 1000 and 5000 observations.

We report Monte Carlo biases, standard deviations, and root mean squared errors for θ using our method in the first three columns of Table 2.1. For the cumulative distribution functions of log(price), we compute the integrated squared biases and integrated mean squared errors, as shown in the first two columns of Table 2.2. These results are based on a sample size of $N = 1000$. The results for $N = 5000$ are provided in Tables 2.3–2.4 in Section 2.10. Our estimator performs well in finite samples across all DGPs we consider. The biases of our estimator are small, and the standard deviation decreases as the sample size increases in all simulation designs.

Table 2.1: Simulation Results for Utility Parameters: $N = 1000$

	Functional Contraction			Two-Step Method		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0075	0.1958	0.1957	0.0027	0.2126	0.2124
ξ_2	0.0021	0.0721	0.0721	0.0003	0.0980	0.0979
β	-0.0010	0.0906	0.0905	0.0016	0.0929	0.0929
			DGP 2			
γ	-0.0087	0.1990	0.1990	0.0196	0.2451	0.2457
ξ_2	0.0021	0.0728	0.0728	0.0183	0.1203	0.1215
β	-0.0049	0.0945	0.0946	0.0036	0.0960	0.0960
			DGP 3			
γ	-0.0254	0.1603	0.1621	0.1704	0.2398	0.2940
ξ_2	-0.0006	0.0702	0.0701	0.0097	0.0860	0.0864
β	-0.0045	0.0930	0.0930	-0.0023	0.0947	0.0946
			DGP 4			
γ	-0.0131	0.3485	0.3484	0.0368	0.3826	0.3840
ξ_2	-0.0016	0.0677	0.0676	-0.0044	0.0731	0.0731
β	-0.0045	0.0933	0.0933	-0.0040	0.0941	0.0941
			DGP 5			
γ	0.0551	0.9650	0.9656	0.1873	0.7830	0.8044
ξ_2	-0.0023	0.0671	0.0671	0.0047	0.0675	0.0676
β	-0.0050	0.0886	0.0886	-0.0050	0.0885	0.0886

Compared to the classic two-step method, our estimator outperforms the standard approach in DGPs 2–5. Because our method allows for nonparametric estimation of the offered price distributions, while the standard method misspecifies the pricing equation, we achieve significantly lower integrated squared bias and mean squared error for the cumulative distribution functions of log(price). This result can also be

Table 2.2: Simulation Results for CDF of log(Price): $N = 1000$

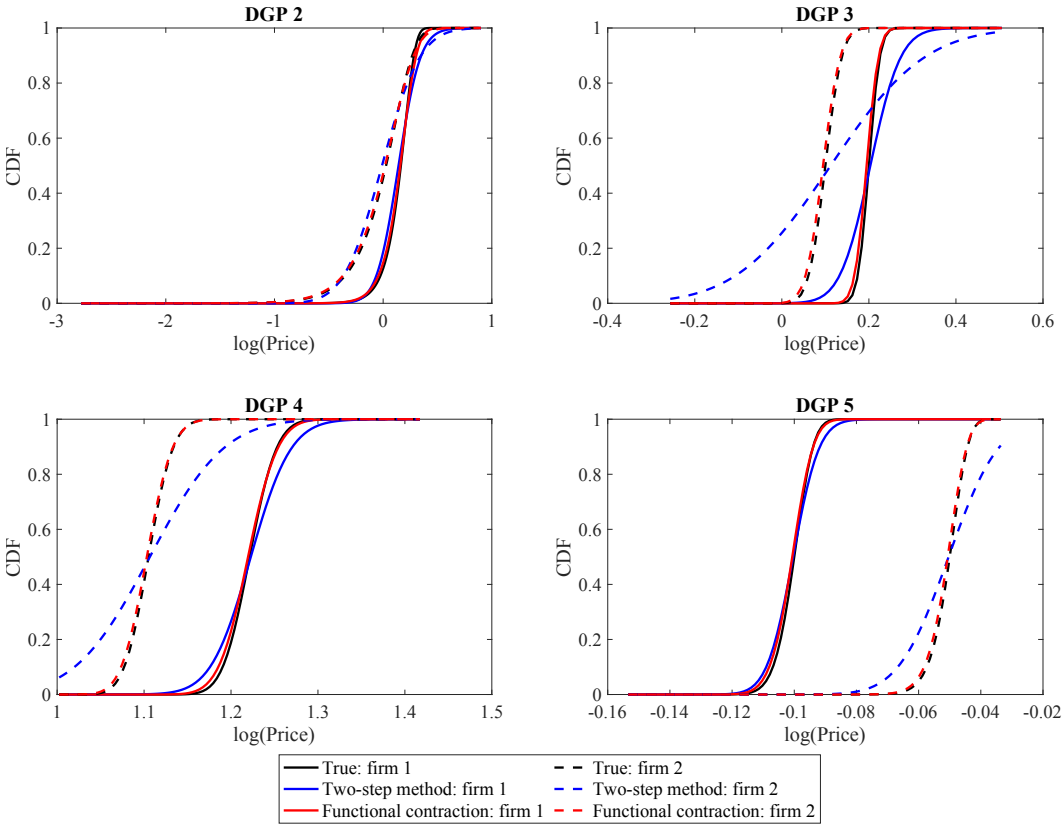
	Func. Contraction		Two-Step Method	
	IBias^2	IMSE	IBias^2	IMSE
	DGP 1			
$F_1(\cdot\|x_{i2}=0)$	0.0003	0.0029	0.0006	0.0211
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0006	0.0000	0.0016
$F_1(\cdot\|x_{i2}=1)$	0.0004	0.0032	0.0003	0.0125
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0013	0.0001	0.0042
	DGP 2			
$F_1(\cdot\|x_{i2}=0)$	0.0006	0.0032	0.0042	0.0269
$F_2(\cdot\|x_{i2}=0)$	0.0002	0.0006	0.0021	0.0040
$F_1(\cdot\|x_{i2}=1)$	0.0008	0.0037	0.0035	0.0177
$F_2(\cdot\|x_{i2}=1)$	0.0003	0.0014	0.0022	0.0070
	DGP 3			
$F_1(\cdot\|x_{i2}=0)$	0.0060	0.0086	0.0247	0.0501
$F_2(\cdot\|x_{i2}=0)$	0.0028	0.0032	0.0499	0.0525
$F_1(\cdot\|x_{i2}=1)$	0.0007	0.0033	0.0022	0.0119
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0013	0.0129	0.0170
	DGP 4			
$F_1(\cdot\|x_{i2}=0)$	0.0007	0.0033	0.0049	0.0304
$F_2(\cdot\|x_{i2}=0)$	0.0008	0.0012	0.0281	0.0303
$F_1(\cdot\|x_{i2}=1)$	0.0005	0.0046	0.0005	0.0161
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0011	0.0087	0.0112
	DGP 5			
$F_1(\cdot\|x_{i2}=0)$	0.0014	0.0034	0.0026	0.0226
$F_2(\cdot\|x_{i2}=0)$	0.0014	0.0018	0.0211	0.0234
$F_1(\cdot\|x_{i2}=1)$	0.0008	0.0058	0.0008	0.0192
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0011	0.0071	0.0086

Note: The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R}\hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

visualized in Figure 2.1, where we plot the true CDFs of log(price) for firms 1 and 2, alongside those obtained using our method and the two-step method.

For the two-step method, the misspecification of the pricing equation also creates a severe bias in estimating the parameters in the selection function. In particular, when the error term exhibits heteroskedasticity (DGP 3) or is nonseparable in the pricing equation (DGPs 4–5), the bias for the price sensitivity parameter γ is large and does not vanish as the sample size increases.

Another key advantage of our approach is that it does not require an instrument to exogenously shift the selection probability. It is well known in the literature that the two-step method is nearly unidentified when the same regressors are used



Note: The black, blue, and red curves represent the true CDF, the CDF estimated using the two-step method, and the CDF estimated using the functional contraction method, respectively. Solid lines represent the CDFs for firm 1, while dashed lines represent those for firm 2.

Figure 2.1: CDF of log(price) for firms 1 and 2 (conditional on $x = 0$).

in both the selection function and the outcome equation. This occurs because the inverse Mills ratio is approximately linear over a wide range of its argument. In our simulations, when the regressor in the pricing equation is discrete, the bias correction term becomes perfectly collinear with the regressor, rendering the two-step method infeasible without an excluded variable in the selection equation.

In contrast, our approach does not require an excluded variable in the selection equation. To illustrate this, we conduct a set of Monte Carlo simulations where the excluded variable x_{i1} is removed from the indirect utility, using the same five DGPs for log(price). The results for this specification are reported in Tables 2.5–2.6 in Section 2.10. As shown, our estimator performs well in finite samples, even without an additional excluded variable to exogenously shift the selection probability. Our estimator consistently shows low bias across different DGPs and exhibits a decreasing standard deviation as the sample size increases.

Our method requires that the functional form of the selection function is known to econometricians. To assess the performance of our estimator when the selection function is misspecified, we conduct a series of Monte Carlo simulations. Specifically, we consider a scenario where the econometrician assumes that ε follows a logistic distribution, while it is actually generated from a normal distribution. In Tables 2.7–2.8 in Section 2.10, we report the estimation results for the utility parameters and CDFs of log(price) under this misspecification. Although we observe a 7–8% bias in the utility parameters, our estimator for the offered price distributions performs well. The integrated squared bias and mean squared errors of the CDFs remain close to those in Table 2.2. This exercise suggests that our estimator for the offered price distributions is robust to misspecification of the selection function, a valuable feature in practice, especially when the econometrician lacks prior knowledge about the form of the selection function.

Finally, we briefly discuss how our functional contraction performs in practice. We compute the modulus ρ^* across all five simulation designs. Except for DGP 2—where the error term in the pricing equations is drawn from extreme value distributions, resulting in a wider price range—the modulus in all other cases is quite small (for example, $\rho^* = 0.37$ in DGP 1).[12] Consequently, our iteration process converges within 3–5 iterations. For DGP 2, although the modulus exceeds 1 ($\rho^* = 1.23$), the iteration process still performs well and converges to the same fixed point, even with different initial values. This is not surprising, as Theorems 3 and 4 provide only sufficient conditions for the contraction.

2.6 Applications

Our estimator introduced in Section 2.4 is broadly applicable to a variety of empirical settings. It effectively addresses the challenge of selection bias, which arises when only the outcomes of chosen alternatives are observed in the data. We impose no parametric restrictions on the potential outcome distribution and allow it to vary flexibly across alternatives. Moreover, the selection function in our model can incorporate alternative-specific unobserved heterogeneity and does not require an excluded variable, which is desirable in many empirical settings. In the following section, we discuss three types of empirical applications: consumer demand

[12] The magnitude of the modulus depends heavily on the product of the price sensitivity parameter γ and price range. In our Monte Carlo simulations, γ is normalized to be 1. In DGP 1, a price range of approximately 2.5 leads to a small $\rho^* = 0.37$. In empirical applications, the price sensitivity parameter γ is around 10^{-3} (for example, see Cosconati et al., 2024). Then with a price range around 2500 euros, the modulus remains small.

estimation, auctions with missing bids, and Roy models.

Consumer Demand

The first application of our method is the standard differentiated product demand estimation pioneered by S. T. Berry (1994) and S. Berry, Levinsohn, and Pakes (1995). In classic demand models, the price of a product is often assumed to be uniform across all consumers (e.g., the list price of a vehicle). But this assumption does not hold in contexts involving price discrimination or personalized pricing (Sagl, 2023; Buchholz et al., 2020; Dubé and Misra, 2023), discount negotiation (Goldberg, 1996; Allen, Clark, and Houde, 2014), or risk-based pricing (Crawford, Pavanini, and Schivardi, 2018; Cosconati et al., 2024). In these contexts, researchers can relatively easily gather data on the transaction prices consumers pay, but it is challenging to gain access to competing prices offered to consumers.

In a companion paper with coauthors (Cosconati et al., 2024), we apply our method to estimate demand and insurance companies' information technology in the auto insurance market, where only the transaction prices of selected insurance plans are observed. In this market, insurance companies employ risk-based pricing. For each consumer, an insurance company generates a noisy estimate of their risk type and prices accordingly. Our goal is to quantify the heterogeneity in insurers' information technology, as measured by the dispersion of their risk estimates. Since the shape of the offered price distribution reflects the distribution of risk estimates, allowing for flexible estimation of the offered price distribution is crucial.

We nonparametrically estimate each insurance company's offered price distribution using our functional contraction approach. In this application, we assume that the offered prices across different firms are independent, conditional on the consumer's true risk type, which is estimated using a panel of *ex-post* realized claim records over multiple years.

In Figure 2.2, we plot the nonparamtrically estimated density functions for prices from several firms. These distributions vary significantly, with noticeable differences in mean, variance, and skewness, across firms, suggesting substantial heterogeneity in their information technology and pricing strategies. Building on this estimation, we further estimate the price sensitivity parameter, firm-specific unobserved heterogeneity (e.g., service quality or brand loyalty), and each firm's information precision. Our findings provides key insight for analyzing competition under various forms of supply-side heterogeneity in this market (Cosconati et al.,

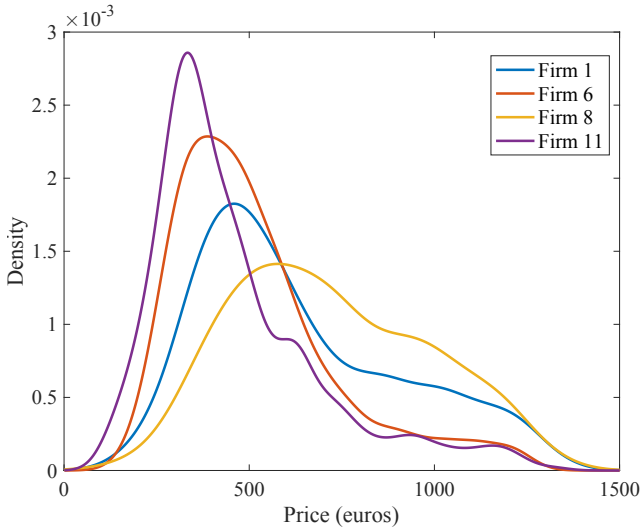


Figure 2.2: Estimated density functions

2024).

From a practical point of view, our iterative procedure to numerically solve for the offered price distributions given demand parameters is easy to implement and performs well in practice. In our empirical application using data from 11 insurers, the iterative algorithm converges very quickly, typically requiring only 6–7 iterations.

Auctions with Missing Bids

In certain auctions, not all bids are available, either due to the auction's structure or incomplete data. For instance, in Dutch auctions, only the winning bid is recorded, as the auction concludes as soon as the first bid is placed. Allen, Clark, Hickman, et al. (2024) study FDIC auctions for insolvent banks, where only the winning and the second-highest bids are recorded. Similarly, U.S. Forest Service timber auctions record only the top fourteen bids, while the Washington State Department of Transportation publishes only the three lowest bids for their highway procurement auctions.

The existing literature has shown that certain types of auction models can be identified using only winning bids or transaction prices. For example, Athey and Haile (2002) show that the symmetric IPV models are identified with the transaction price by exploiting a one-to-one mapping between an order statistic and its parent distribution. Komarova (2013) analyzes asymmetric second-price auctions where only the winning bids and the winner's identity are observed. A related result for generalized competing risks models can be found in Meilijson (1981). More recently,

Guerre and Y. Luo (2019) examine nonparametric identification of symmetric IPV first-price auctions with only winning bids, accounting for unobserved competition.

Our method is valuable for nonparametrically recovering the complete bid distribution and the auctioneer's scoring weights in multi-attribute auctions when the data contain only the winning bids and winner's identity, particularly in the presence of bidder asymmetry.[13] Auctions in many settings have used the scoring rule that departs from the lowest bid criterion by accounting for quality differences (Asker and Cantillon, 2008; Lewis and Bajari, 2011; Nakabayashi, 2013; Yoganarasimhan, 2016; Takahashi, 2018; Krasnokutskaya, Song, and Tang, 2020; Allen, Clark, Hickman, et al., 2024). Our selection model is closely related to Krasnokutskaya, Song, and Tang (2020), which employs a discrete choice framework with unknown, buyer-specific weights in the scoring rule. We allow the scoring rule to depend on both observed (x_{ij}) and unobserved bidder heterogeneity (ξ_j), with the error term (ε_{ij}) capturing uncertainty in the scoring rule.[14]

Beyond independent private value models, our method can be applied to certain common value auction models, such as the mineral rights model, where bidders' signals are assumed to be independent conditional on the common value. In these auctions, we can recover the bid distributions conditional on the ex-post realized common value.

Roy Models

Another important application of our method is estimating Roy models (Roy, 1951) in labor market contexts. Variants of the Roy model have been widely used in the literature to study decisions such as whether to continue schooling (Willis and Rosen, 1979), which occupation to pursue (Heckman and Sedlacek, 1985), whether to join a union (L.-F. Lee, 1978), and whether to migrate (Borjas, 1987). Our selection model falls within the framework of "Generalized Roy Model", as defined by Heckman and Vytlacil (2007). We allow the utility that individual i gains from alternative j to depend not only on prices (or wages in labor market contexts) but also on non-pecuniary aspects of the alternative, either observable or unobservable

[13] Flexibly accommodating bidder asymmetries is known to be challenging in auction models (see discussions in the handbook chapter by Athey and Haile, 2007). Bidder asymmetries may arise from factors such as distance to the contract location (Flambard and Perrigne, 2006), information advantages (Hendricks and Porter, 1988; De Silva, Kosmopoulou, and Lamarche, 2009), varying risk attitudes (Campo, 2012), or strategic sophistication (Hortaçsu et al., 2019).

[14] Other recent papers that consider unknown weights in the scoring rule include Takahashi (2018) and Allen, Clark, Hickman, et al. (2024).

to the econometrician. The comparison between our approach and standard two-step methods for estimating Roy models has already been discussed in the introduction; therefore, we do not reiterate it here.

2.7 Conclusion

We introduce a novel method for estimating nonseparable selection models when only a selected sample of outcomes is observed. We show that potential outcome distributions can be nonparametrically identified from the observed distribution of selected outcomes, given a selection function. We achieve this by constructing an operator whose fixed point represents the potential outcome distributions and proving that this operator is a functional contraction. Building on this theoretical result, we propose a two-step semiparametric maximum likelihood estimator for both the selection function and potential outcome distributions. The consistency and asymptotic normality of the proposed estimator are established.

Our approach fundamentally differs from the classic two-step method for addressing sample selection bias. We allow the outcome equation to be fully nonparametric and nonseparable in error terms. Our goal is to recover the entire distribution of potential outcomes rather than focusing on specific moments or quantiles. In essence, we correct for sample selection bias across the entire distribution of potential outcomes by examining how the bias is *systematically* generated by the selection model. This approach allows for fully heterogeneous effects of covariates on outcomes, which is a crucial feature for empirical analysis, as discussed in Chernozhukov, Iván Fernández-Val, and S. Luo, 2023. Another key advantage of our approach is that it does not rely on instruments to exogenously shift selection probabilities, which are often challenging to find in empirical settings, or on identification-at-infinity arguments. Our approach also accommodates asymmetry in outcome distributions across alternatives and flexibly incorporates unobserved alternative-specific heterogeneity in the selection model.

We find that the proposed estimation strategy performs well in both simulations and real-world data applications (see our demand estimation using insurance market data in Cosconati et al., 2024). Moreover, our approach is straightforward to implement and computationally efficient, making it highly appealing to empirical researchers. The estimator can be readily applied to a variety of empirical settings where only a selected sample of outcomes is observed, including consumer demand models with only transaction prices, auctions with incomplete bid data, and various selection

models in labor economics. Our method is particularly valuable in applications where the entire distribution of outcomes is of interest.

2.8 Connection to Quantal Response Equilibria

In this section, we connect our result to the quantal response equilibria (McKelvey and Palfrey, 1995).

Let us rename our variables. There is a set $\mathcal{J} = \{1, 2, \cdots, J\}$ of players. For each player $j \in \mathcal{J}$, there is a finite set $P_j = \{p_{j1}, p_{j2}, \cdots, p_{jn_j}\} \subset [\underline{p}_j, \overline{p}_j]$ consisting of n_j pure strategies. A payoff function $f\colon \prod_{j\in\mathcal{J}} P_j \to \Delta(\mathcal{J})$ assigns payoff f_j to player j. Let $g_j \in \Delta P_j$ denote player j' mixed strategy and $g = \prod_{j\in\mathcal{J}} g_j$. The player j's expected payoff for playing pure strategy p_j, given other players' strategy g_{-j}, is

$$Pr_j(p_j; g) = \int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k\neq j} g_k(p_k).$$

We define the quantal response operator $\mathbb{T}\colon \prod_j \Delta(P_j) \to \prod_j \Delta(P_j)$ by

$$(\mathbb{T}g)_j(p_j) = \frac{\exp(-\lambda Pr_j(p_j; g))}{\sum_{p_j\in P_j} \exp(-\lambda Pr_j(p_j; g))}.$$

In words, given the expected payoff $Pr_j(p_j; g)$, player j's probability of playing strategy p_j is proportional to $\exp(-\lambda Pr_j(p_j; g))$. Lemma 1 in McKelvey and Palfrey, 1995 states that operator $\mathbb{T}$ is a contraction for a sufficiently small λ. This is intuitive as $\mathbb{T}$ sends probability measures to the center of the simplex when λ is small.

Note that our operator T is quite different. By definition,

$$(T\Psi)_j(p) = \frac{\int_{\underline{p}_j}^{p} d\tilde{G}_j(y)/Pr_j(y; \Psi)}{\int_{\underline{p}_j}^{\overline{p}_j} d\tilde{G}_j(y)/Pr_j(y; \Psi)}.$$

Given the expected probability Pr, to compute the new measure, each p_j is weighted by $d\tilde{G}(p_j)$, where $\tilde{G}$ can be any measure. This distinction complicates our problem. With the sup norm, McKelvey and Palfrey, 1995 show that $\mathbb{T}$ is a contraction for sufficiently small λ. However, the presence of $\tilde{G}$ renders the sup norm not suitable for our task. Instead, our metric d is designed specifically to deal with $\tilde{G}$.

2.9 Omitted Proofs

Proof of Theorem 3

Lemma 9. *For two probability measures $S, Q \in \Delta(Y)$, $\delta > 0$,*

$$\sup_{d(S,Q)\leq\delta} ||S - Q||_{TV} \leq \delta/2.$$

Proof of Lemma 9. We first consider the case where Y contains only two elements. Then we can identify S with $(p, 1)$ for some $p \in [0, 1]$. We can pin down the Q that achieves the maximum $||S - Q||_{TV}$ under the constraint that $d(S, Q) \leq \delta$. At the maximum, this constraint is binding. Let $Q = (p - \epsilon, 1)$. By $d(S, Q) = \delta$,

$$\ln \frac{p}{p-\epsilon} + \ln \frac{1-p+\epsilon}{1-p} = \delta. \tag{2.13}$$

We can solve for ϵ

$$\epsilon = \frac{p(1-p)(e^\delta - 1)}{p + (1-p)e^\delta}.$$

Plug this into the total variation norm

$$\frac{1}{2}||S - Q||_{TV} = \epsilon = (e^\delta - 1)[\frac{1}{1-p} + \frac{e^\delta}{p}]^{-1}.$$

Then we take sup over p. Note that $\frac{1}{1-p} + \frac{e^\delta}{p}$ as a function of p is convex and achieves a unique minimum at $p = \frac{e^{\delta/2}}{1+e^{\delta/2}}$. As a result,

$$\sup_{d(S,Q)\leq\delta} \frac{1}{2}||S - Q||_{TV} = \frac{(e^\delta - 1)}{(1 + e^{\delta/2})^2} = \frac{e^{\delta/2} - 1}{e^{\delta/2} + 1}.$$

To show $\sup_{d(S,Q)\leq\delta} ||S - Q||_{TV} \leq \delta/2$, it suffices to show that for all $\delta \geq 0$,

$$\frac{e^{\delta/2} - 1}{e^{\delta/2} + 1} \leq \delta/4$$

which holds true.[15] Note that the limiting case $\delta \to 0$, $p = \frac{1}{2}$, $\epsilon = \frac{\delta}{4}$ achieves this upper bound.

[15] To see this,

$$\begin{aligned} &\frac{e^\delta - 1}{e^\delta + 1} \leq \delta/2 \\ \Leftrightarrow &1 - \frac{2}{e^\delta + 1} \leq \frac{\delta}{2} \\ \Leftrightarrow &2 - \delta \leq \frac{4}{e^\delta + 1} \end{aligned}$$

which is true since function $\frac{4}{e^\delta+1}$ is convex and is tangent to the function $2 - \delta$ at $\delta = 0$.

Now we prove this lemma for a general space Y and general CDF. For any $S, Q \in \Delta(Y)$ and $d(S,Q) \leq \delta$. Define two functions

$$P_Q(S,Q) = \int_{y \in Y: \frac{dS}{dQ}(y) \geq 1} dQ(y)$$

$$P_S(S,Q) = \int_{y \in Y: \frac{dS}{dQ}(y) \geq 1} dS(y).$$

Note that

$$\frac{P_S(S,Q)}{P_Q(S,Q)} \leq \operatorname*{ess\,sup}_{y \in Y} \frac{dS}{dQ}(y)$$

$$\frac{1 - P_Q(S,Q)}{1 - P_S(S,Q)} \leq \operatorname*{ess\,sup}_{y \in Y} \frac{dQ}{dS}(y)$$

which implies

$$\ln \frac{P_S(S,Q)}{P_Q(S,Q)} + \ln \frac{1 - P_Q(S,Q)}{1 - P_S(S,Q)} \leq \operatorname{ess\,sup} \ln \frac{dS}{dQ}(y) + \operatorname{ess\,sup} \ln \frac{dQ}{dS}(y) \leq \delta$$

since $d(S,Q) \leq \delta$. Observe that here $P_S(S,Q)$ faces the same constraint as p in the two-point support case in Equation (2.13). Thus, the total variation norm

$$||S - Q||_{TV} = 2[P_S(S,Q) - P_Q(S,Q)] \leq \delta/2.$$

□

Proof of Theorem 3. Recall that

$$Pr_j(p_j; \Psi) = \int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k, k \neq j} d\Psi_k(p_k).$$

Define the ratio function

$$R_j(p_j; \Psi, \Phi) = \frac{Pr_j(p_j; \Psi)}{Pr_j(p_j; \Phi)}.$$

We show that for all $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$D(T\Psi, T\Phi) \leq \rho D(\Psi, \Phi).$$

Given Equation (2.7) and the definition of the metric d, we have

$$d((T\Psi)_j, (T\Phi)_j) \leq \sup_{p_j} \ln R_j(p_j; \Psi, \Phi) - \inf_{p_j} \ln R_j(p_j; \Psi, \Phi).$$

The equality holds when $\tilde{G}_j$ admits full support on $[\underline{p}_j, \overline{p}_j]$. Thus, it suffices to show that for all $j \in \mathcal{J}$

$$\sup_{p_j} \ln R_j(p_j; \Psi, \Phi) - \inf_{p_j} \ln R_j(p_j; \Psi, \Phi) \leq \rho D(\Psi, \Phi) \tag{2.14}$$

We evaluate how the log ratio changes with p_j,

$$\frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} = \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)} - \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} \prod_{k,k\neq j} d\Phi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Phi_k(p_k)} \tag{2.15}$$

$$= \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} f_j \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)} - \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} f_j \prod_{k,k\neq j} d\Phi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Phi_k(p_k)}. \tag{2.16}$$

Next, we define a new measure $f_j\Psi_{-j} \in \Delta(\prod_{k\neq j}[\underline{p}_k, \overline{p}_k])$

$$f_j\Psi_{-j}(y) = \frac{\int_{\underline{\boldsymbol{p}}_{-j}}^{y} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)}.$$

Similarly, we define measure $f_j\Phi_{-j} \in \Delta(\prod_{k\neq j}[\underline{p}_k, \overline{p}_k])$. (Both measures depend on p_j.) Given these measures, we can rewrite Equation (2.16)

$$\frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} = \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j} \sim f_j\Psi_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} - \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j} \sim f_j\Phi_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} \tag{2.17}$$

$$= \int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} [df_j\Psi_{-j}(\boldsymbol{p}_{-j}) - df_j\Phi_{-j}(\boldsymbol{p}_{-j})]. \tag{2.18}$$

We shall upper bound this integral under the constraint $D(\Psi, \Phi) \leq \delta$ for some arbitrary $\delta > 0$.

$$\sup_{D(\Psi,\Phi)\leq\delta} \left| \frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} \right| = \sup_{D(\Psi,\Phi)\leq\delta} \left| \int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} [df_j\Psi_{-j}(\boldsymbol{p}_{-j}) - df_j\Phi_{-j}(\boldsymbol{p}_{-j})] \right|$$
$$\leq M_j \sup_{D(\Psi,\Phi)\leq\delta} \frac{1}{2} ||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV}.$$

The inequality follows by interpreting the integral as a transportation problem. We transport the mass from distribution $f_j\Phi_{-j}$ to $f_j\Psi_{-j}$. The function $\frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j}$ is

the height. Then the integral is the change in the gravitational potential, which is bounded by the product of the total transportation mass $\frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV}$ and the largest height difference, M_j. Note that given $D(\Psi, \Phi) \leq \delta$,

$$d(f_j\Psi_{-j}, f_j\Phi_{-j}) = d(\Psi_{-j}, \Phi_{-j}) \leq (J-1)\delta,$$

as for all j, $d(\Psi_j, \Phi_j) \leq D(\Psi, \Phi) \leq \delta$. Thus, for all $\delta > 0$,

$$\begin{aligned}\sup_{D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\right| \leq & M_j \sup_{D(\Psi,\Phi)\leq\delta}\frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV} \\ \leq & M_j \sup_{d(f_j\Psi_{-j},f_j\Phi_{-j})\leq(J-1)\delta}\frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV} \\ \leq & M_j\frac{1}{4}(J-1)\delta, \end{aligned} \tag{2.19}$$

where the last inequality follows by Lemma 9. By Lemma 10,

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| = \sup_{D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}M_j.$$

To see why the inequality holds, towards a contradiction, suppose it does not hold. Then there exists $\tilde{\Psi}$, $\tilde{\Phi}$ with $D(\tilde{\Psi}, \tilde{\Phi}) = \delta_1$ and

$$\left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\frac{1}{D(\tilde{\Psi},\tilde{\Phi})}\right| > \frac{J-1}{4}M_j$$

$$\left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\right| > \frac{J-1}{4}M_jD(\tilde{\Psi},\tilde{\Phi})$$

which implies that

$$\sup_{D(\Psi,\Phi)\leq\delta_1}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \geq \left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\frac{1}{D(\tilde{\Psi},\tilde{\Phi})}\right| > \frac{J-1}{4}M_j$$

contradicting Equation (2.19) which holds for all $\delta > 0$.

By the fundamental theorem of calculus, for all $p_j, p_j' \in [\underline{p}_j, \overline{p}_j]$,

$$\sup_{\Psi,\Phi}\left|\frac{\ln R_j(p_j;\Psi,\Phi) - \ln R_j(p_j';\Psi,\Phi)}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}M_j(\overline{p}_j - \underline{p}_j).$$

Finally, for all $j \in \mathcal{J}$, all Ψ, Φ,

$$\sup_{p_j}\ln R_j(p_j;\Psi,\Phi) - \inf_{p_j}\ln R_j(p_j;\Psi,\Phi) \leq \rho D(\Psi,\Phi).$$

□

Lemma 10. *For all $\delta > 0$,*

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| = \sup_{\Psi,\Phi,D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right|. \tag{2.20}$$

Proof of Lemma 10. We prove this lemma through a continuous interpolation. Fixing any $\Psi,\Phi \in \prod_j \Delta([\underline{p}_j,\overline{p}_j])$, we define a continuous interpolation $\Upsilon(\cdot;\lambda) \in \prod_j \Delta([\underline{p}_j,\overline{p}_j])$ parametrized by $\lambda \in [0,1]$:

$$\Upsilon_j(p_j;\lambda) = \frac{\int_{\underline{p}_j}^{p_j} d\Phi_j(y)\cdot\left(\frac{d\Psi_j}{d\Phi_j}(y)\right)^\lambda}{\int_{\underline{p}_j}^{\overline{p}_j} d\Phi_j(y)\cdot\left(\frac{d\Psi_j}{d\Phi_j}(y)\right)^\lambda}$$

Notice that $\Upsilon(\cdot;0)=\Phi$, $\Upsilon(\cdot;1)=\Psi$. Moreover,

$$d(\Upsilon_j(\cdot;\lambda_1),\Upsilon_j(\cdot;\lambda_2)) = |\lambda_1-\lambda_2|d(\Psi_j,\Phi_j).$$

Thus, in our metric space, $\Upsilon(\cdot;\lambda)$ is an interpolation that is linear in the metric.[16] That is, for all $\lambda_1,\lambda_2 \in [0,1]$,

$$D(\Upsilon(\cdot;\lambda_1),\Upsilon(\cdot;\lambda_2)) = |\lambda_1-\lambda_2|D(\Psi,\Phi).$$

We define a new function by adapting Equation (2.17).

$$k(\lambda) = \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j}\sim f_j\Upsilon_{-j}(\cdot;\lambda)}\frac{\partial\ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} - \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j}\sim f_j\Phi_{-j}}\frac{\partial\ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j}.$$

Notice that when $\lambda = 1$, this reduces to Equation (2.17). As k is continuously differentiable, there exists $0 \leq \underline{\lambda} < \underline{\lambda} + d\lambda \leq 1$ and $d\lambda \leq \frac{\delta}{D(\Psi,\Phi)}$ such that

$$|k(1)| \leq \left|\frac{k(\underline{\lambda}+d\lambda)-k(\underline{\lambda})}{d\lambda}\right|$$

[16] Note that $\Upsilon(\cdot;\lambda)$ is also a linear interpolation in the Kullback-Leibler divergence, since

$$D_{KL}(\Phi||\Upsilon(\cdot;\lambda)) = \lambda D_{KL}(\Phi||\Psi)$$

and

$$D_{KL}(\Psi||\Upsilon(\cdot;\lambda)) = (1-\lambda)D_{KL}(\Psi||\Phi).$$

This is equivalent to

$$\left|\frac{k(1)}{D(\Psi,\Phi)}\right| \leq \left|\frac{k(\underline{\lambda}+d\lambda)-k(\underline{\lambda})}{d\lambda D(\Psi,\Phi)}\right|$$
$$\Leftrightarrow \left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \left|\frac{d\ln R_j(p_j;\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}{dp_j}\frac{1}{d\lambda D(\Psi,\Phi)}\right|$$
$$\Leftrightarrow \left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \left|\frac{d\ln R_j(p_j;\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}{dp_j}\frac{1}{D(\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}\right|$$

As $D(\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda})) = d\lambda D(\Psi,\Phi) \leq \delta$, we have established Equation (2.20). □

Proof of Theorem 4

Proof of Theorem 4. With Assumption 1, we can provide tighter bound on the right-hand side of Equation (2.18).

$$\begin{aligned}
&\sup_{D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\right| \\
=&\sup_{D(\Psi,\Phi)\leq\delta}\left|\int_{\boldsymbol{p}_{-j}}\frac{\partial\ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j}[df_j\Psi_{-j}(\boldsymbol{p}_{-j})-df_j\Phi_{-j}(\boldsymbol{p}_{-j})]\right| \\
\leq&\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right]\sup_{D(\Psi,\Phi)\leq\delta}\frac{1}{2}||f_j\Psi_{-j}-f_j\Phi_{-j}||_{TV} \\
\leq&\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right]\frac{J-1}{4}\delta.
\end{aligned}$$

By Lemma 10,

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right].$$

By the fundamental theorem of calculus, for all $p_j, p_j' \in [\underline{p}_j, \overline{p}_j]$,

$$\sup_{\Psi,\Phi}\left|\frac{\ln R_j(p_j;\Psi,\Phi)-\ln R_j(p_j';\Psi,\Phi)}{D(\Psi,\Phi)}\right| \leq \rho^*.$$

Finally, for all $j \in \mathcal{J}$, all Ψ, Φ,

$$\sup_{p_j}\ln R_j(p_j;\Psi,\Phi)-\inf_{p_j}\ln R_j(p_j;\Psi,\Phi) \leq \rho^* D(\Psi,\Phi).$$

□

Proof of Theorem 5

Proof of Proposition 5. Suppose $\rho < 1$. By Theorem 3, the operator T is a contraction. This implies that F is surjective, since for any $\tilde{G}$, we can take a $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$F(\lim_{n\to\infty} T^n \Psi) = \tilde{G}.$$

Moreover, F is injective. Towards a contradiction, suppose F maps both $G_1 \neq G_2 \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ to the same $\tilde{G}$. Then both G_1 and G_2 are fixed points for operator T, contradicting contraction.

The mapping F is continuous by Equation (2.1) and (2.2). Take two offered distributions G and G'. By Equation (2.2) and the definition of our metric,

$$\begin{aligned}
d(F(G)_j, F(G')_j) &= \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{dG_j}{dG'_j}(p)\frac{Pr_j(p;G)}{Pr_j(p;G')}\right) + \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{dG'_j}{dG_j}(p)\frac{Pr_j(p;G')}{Pr_j(p;G)}\right) \\
&\leq \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \frac{dG_j}{dG'_j}(p) + \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \frac{dG'_j}{dG_j}(p) \\
&+ \ln \sup_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{Pr_j(p;G)}{Pr_j(p;G')}\right) + \ln \sup_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{Pr_j(p;G')}{Pr_j(p;G)}\right) \\
&\leq D(G,G') + \rho D(G,G'),
\end{aligned}$$

where the last inequality is by Equation (2.14). Consequently,

$$D(F(G), F(G')) \leq (1+\rho) D(G,G'),$$

F is Lipschitz continuous with Lipschitz constant $1+\rho$.

Next, we show F^{-1} is Lipschitz continuous. Take two selected distributions $\tilde{G} \neq \tilde{G}' \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ where $\tilde{G} = F(G)$. Let $T_{\tilde{G}}$ and $T_{\tilde{G}'}$ denote the corresponding operator T. Here we express dependence on the selected distribution. Note that

$$D(\tilde{G}, \tilde{G}') = D(T_{\tilde{G}}G, T_{\tilde{G}'}G) = D(G, T_{\tilde{G}'}G),$$

where the first equality is by the definition of the operator T and the metric D, while the second equality is by G being a fixed point of $T_{\tilde{G}}$. Observe that

$$D(T^k_{\tilde{G}'}G, T^{k+1}_{\tilde{G}'}G) \leq \rho^k D(G, T_{\tilde{G}'}G) = \rho^k D(\tilde{G}, \tilde{G}')$$

$$\begin{aligned} D(F^{-1}(\tilde{G}), F^{-1}(\tilde{G}')) = D(G, F^{-1}(\tilde{G}')) =& D(G, T^{\infty}_{\tilde{G}'}G) \\ \leq& \sum_{k=0}^{\infty} D(T^{k}_{\tilde{G}'}G, T^{k+1}_{\tilde{G}'}G) \\ \leq& \sum_{k=0}^{\infty} \rho^{k} D(\tilde{G}, \tilde{G}') \\ =& \frac{1}{1-\rho} D(\tilde{G}, \tilde{G}'), \end{aligned}$$

where the first inequality is by triangular inequality. This proves that F^{-1} is Lipschitz continuous with Lipschitz constant $\frac{1}{1-\rho}$. □

For proofs below, it suffices to prove the case without variable x. So we shall drop it. We next prove the consistency result (Proposition 5). The proof requires a combination of Lemma 11-14 below. We first collect useful notations below. Let

$$Q_0(\theta) = \sum_j \ln\left(Prob^*_j(\theta, \tilde{G})\right) \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta_0) dG(\boldsymbol{p}).$$

$$\hat{Q}^*_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\left(Prob^*_j(\theta, \hat{G})\right),$$

$$Prob^*_j(\theta, \hat{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta) dF^{-1}(\hat{G}, \theta)(\boldsymbol{p}).$$

$$\hat{Q}_{n,m}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\left(Prob_j(\theta, \hat{G}, m)\right),$$

$$Prob_j(\theta, \hat{G}, m) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta) d\left(T^{m}_{\hat{G},\theta}\Psi\right)(\boldsymbol{p}).$$

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\left(Prob_j(\theta, \hat{G}, m(n))\right),$$

$$\mathfrak{g}^*(z_i, \theta, \hat{G}) = \nabla_\theta \left(\sum_{j=1}^{J} y_{ij} \ln Prob^*_j(\theta, \hat{G}) \right),$$

$$\mathfrak{g}(z_i, \theta, \hat{G}, n) = \nabla_\theta \left(\sum_{j=1}^{J} y_{ij} \ln Prob_j(\theta, \hat{G}, m(n)) \right).$$

Lemma 11. *$F^{-1}(\hat{G}, \theta)$ is continuous in θ.*

Proof of Lemma 11. Let $\theta, \theta' \in \Theta$. Let

$$\tilde{G} = F(G;\theta)$$
$$G' = F^{-1}(\tilde{G};\theta')$$
$$\tilde{G}' = F(G';\theta).$$

As $\theta' \to \theta$, by $F(G';\theta)$ being continuous in θ, $\tilde{G} \to \tilde{G}'$. By $F^{-1}(\tilde{G};\theta)$ being continuous in $\tilde{G}$ (Proposition 5), $F^{-1}(\tilde{G};\theta) \to F^{-1}(\tilde{G}';\theta)$. This is equivalent to $G' \to G$, which is $F^{-1}(\tilde{G};\theta') \to F^{-1}(\tilde{G};\theta)$. This implies that F^{-1} is continuous in θ.

□

For the next lemma, we view $F^{-1}(\theta;\hat{G})$ as a function of θ parametrized by $\hat{G}$.

Lemma 12. *The function $F^{-1}(\theta;\hat{G})$ is equicontinuous in θ, i.e., for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,*

$$D(F^{-1}(\theta;\hat{G}), F^{-1}(\theta';\hat{G})) \leq \epsilon.$$

Proof of Lemma 12. Since the function f is continuous on a compact set $\prod_j [\underline{p}_j, \overline{p}_j] \times \Theta$ and the image of f is in the interior of the simplex, there exists $\underline{f}$ and $\overline{f}$, $0 < \underline{f} \leq \overline{f} < 1$ such that for all $j \in \mathcal{J}$, $\theta \in \Theta$, $\boldsymbol{p} \in \prod_j [\underline{p}_j, \overline{p}_j]$,

$$\underline{f} < f_j(\boldsymbol{p};\theta) < \overline{f}.$$

Consequently, for all $j \in \mathcal{J}$, $\theta \in \Theta$, $p_j \in [\underline{p}_j, \overline{p}_j]$, $G \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$\underline{f} < Pr_j(p_j;G,\theta) < \overline{f}. \tag{2.21}$$

Moreover, since the function f is continuous on a compact set $\prod_j [\underline{p}_j, \overline{p}_j] \times \Theta$, f is uniformly continuous. Thus, for any $\epsilon' > 0$, there exists a $\delta' > 0$ such that for all $j \in \mathcal{J}$, $\boldsymbol{p} \in \prod_j [\underline{p}_j, \overline{p}_j]$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$|f_j(\boldsymbol{p},\theta) - f_j(\boldsymbol{p},\theta')| < \epsilon'.$$

Therefore, for all $j \in \mathcal{J}$, $p_j \in [\underline{p}_j, \overline{p}_j]$, $G \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$\begin{aligned} &|Pr_j(p_j;G,\theta) - Pr_j(p_j;G,\theta')| \\ =&\left| \int_{\boldsymbol{p}_{-j}} [f_j(p_j, \boldsymbol{p}_{-j};\theta) - f_j(p_j, \boldsymbol{p}_{-j};\theta')] \prod_{k,k\neq j} dG_k(p_k) \right| < \epsilon'. \end{aligned} \tag{2.22}$$

Take an arbitrary $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. Let $G_\theta = F^{-1}(\theta; \hat{G})$, $G_{\theta'} = F^{-1}(\theta'; \hat{G})$. Let T_θ and $T_{\theta'}$ be the operator T associated with selected distribution $\hat{G}$, when the parameter is θ and θ', respectively: for any $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$(T_\theta \Psi)_j(p) = \frac{\int_{\underline{p}_j}^{p} d\hat{G}_j(y)/Pr_j(y; \Psi, \theta)}{\int_{\underline{p}_j}^{\overline{p}_j} d\hat{G}_j(y)/Pr_j(y; \Psi, \theta)}.$$

By the definition of metric D,

$$D(T_\theta G_\theta, T_{\theta'} G_\theta) \leq \max_j \left[\sup_p \ln \frac{Pr_j(p; G_\theta, \theta)}{Pr_j(p; G_\theta, \theta')} + \sup_p \ln \frac{Pr_j(p; G_\theta, \theta')}{Pr_j(p; G_\theta, \theta)} \right].$$

By Equation (2.21) and (2.22), for all $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$D(T_\theta G_\theta, T_{\theta'} G_\theta) \leq 2 \ln \frac{\overline{f} + \epsilon'}{\underline{f}},$$

$$\begin{aligned} D(F^{-1}(\theta; \hat{G}), F^{-1}(\theta'; \hat{G})) =& D(G_\theta, T_{\theta'}^{\infty} G_\theta) \\ \leq& \sum_{k=0}^{\infty} D(T_{\theta'}^{k} G_\theta, T_{\theta'}^{k+1} G_\theta) \\ \leq& \sum_{k=0}^{\infty} \bar{\rho}^k D(G_\theta, T_{\theta'} G_\theta) \\ =& \frac{1}{1-\bar{\rho}} D(T_\theta G_\theta, T_{\theta'} G_\theta) \\ \leq& \frac{2}{1-\bar{\rho}} \ln \frac{\overline{f} + \epsilon'}{\underline{f}}. \end{aligned}$$

Finally, for any $\epsilon > 0$, let ϵ' be such that $\frac{2}{1-\bar{\rho}} \ln \frac{\overline{f}+\epsilon'}{\underline{f}} = \epsilon$. The δ' corresponding to this ϵ' is the desired δ in the statement of the Lemma.

□

Lemma 13. *$\hat{Q}_n^*(\theta)$ converges uniformly in probability to $Q_0(\theta)$.*

Proof of Lemma 13. By Lemma 12 and the uniform continuity of f, for all j, $Prob_j^*(\theta; \hat{G})$ is equicontinuous in θ, parametrized by $\hat{G}$. That is, for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$|Prob_j^*(\theta; \hat{G}) - Prob_j^*(\theta'; \hat{G})| \leq \epsilon.$$

Consequently, by Equation (2.21), for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\{z_i\}_{i=1}^n$,

$$|\hat{Q}_n^*(\theta) - \hat{Q}_n^*(\theta')| \leq \ln \frac{\overline{f} + \epsilon}{\underline{f}}.$$

Thus, $\hat{Q}_n^*(\theta)$ is equicontinuous in θ.

For all $\theta \in \Theta$, $\hat{Q}_n^*(\theta)$ converges in probability to $Q_0(\theta)$, by the weakly law of large numbers, $\hat{G} \xrightarrow{p} \tilde{G}$, and F^{-1} being continuous (Proposition 5). Lastly, $\hat{Q}_n^*(\theta)$ converges uniformly in probability to $Q_0(\theta)$, as $\hat{Q}_n^*$ is equicontinuous in θ (Lemma 2.8 in Newey and McFadden, 1994). □

Lemma 14. *$\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$.*

Proof of Lemma 14. Pick a $\Psi \sim \hat{G}$.[17] Fix some $\epsilon' > 0$. As $\hat{G} \xrightarrow{p} \tilde{G}$, there exists some $\delta(n) \to 0$ as $n \to \infty$ such that

$$D(\hat{G}, \tilde{G}) < \epsilon' \quad \text{with probability above } 1 - \delta(n).$$

Moreover, with probability approaching 1, we have $\hat{G} \sim \tilde{G}$ and thus $\Psi \sim G$. Given $\tilde{G}$, let

$$\overline{D} = \max_{\theta \in \Theta, D(\hat{G}, \tilde{G}) < \epsilon'} D(\Psi, F^{-1}(\hat{G}; \theta)) \leq \max_{\theta \in \Theta} D(\Psi, F^{-1}(\tilde{G}; \theta)) + \frac{\epsilon'}{1 - \bar{\rho}},$$

where the second inequality follows by for all θ, $F^{-1}(\tilde{G}; \theta)$ being Lipschitz continuous in $\tilde{G}$ with Lipschitz constant $\frac{1}{1-\bar{\rho}}$ and the triangle inequality. Note that with probability approaching 1, $\max_{\theta \in \Theta} D(\Psi, F^{-1}(\tilde{G}; \theta))$ is well-defined, since (1). $\Psi \sim \tilde{G}$ with probability approaching 1, (2). $F^{-1}(\tilde{G}; \theta)$ is continuous in θ by Lemma 11, (3). metric D is continuous and Θ is compact.

Next, I show that with probability above $1 - \delta(n)$, $\hat{Q}_{n,m} \to \hat{Q}_n^*$ uniformly in probability as $m \to +\infty$, and the convergence speed does not depend on n. Fix some n. Note

$$Prob_j(\theta, \hat{G}, m) - Prob_j^*(\theta, \hat{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta) d\big(T_{\hat{G},\theta}^m \Psi - F^{-1}(\hat{G}, \theta)\big)(\boldsymbol{p}).$$

With probability above $1 - \delta(n)$, we have $D(\hat{G}, \tilde{G}) < \epsilon'$,

$$D(T_{\hat{G},\theta}^m \Psi, F^{-1}(\hat{G}, \theta)) \leq \bar{\rho}^m D(\Psi, F^{-1}(\hat{G}, \theta)) \leq \bar{\rho}^m \overline{D}$$

[17] Even if Ψ is not equivalent to $\hat{G}$, $T_{\hat{G}}\Psi$ is equivalent to $\hat{G}$.

$$\begin{aligned}\left|Prob_j(\theta,\hat{G},m) - Prob_j^*(\theta,\hat{G})\right| &\leq \left| \sup_{D(\Phi,\Upsilon)\leq\bar{\rho}^m\overline{D}} \int_{\boldsymbol{p}} f_j(\boldsymbol{p};\theta) d\big(\Phi-\Upsilon\big)(\boldsymbol{p})\right| \\ &\leq (\overline{f}-\underline{f})\frac{1}{2} \sup_{D(\Phi,\Upsilon)\leq\bar{\rho}^m\overline{D}} ||\Phi-\Upsilon||_{TV} \\ &\leq (\overline{f}-\underline{f})\frac{1}{2}\frac{1}{2}J\bar{\rho}^m\overline{D},\end{aligned}$$

where the last inequality is by applying Lemma 9 to the product measure. (Here we have an additional factor of J.[18]) Consequently,

$$\left|\hat{Q}_{n,m}(\theta) - \hat{Q}_n^*(\theta)\right| \leq \ln \frac{\underline{f} + \frac{1}{4}(\overline{f}-\underline{f})J\bar{\rho}^m\overline{D}}{\underline{f}} \quad \text{with probability above } 1-\delta(n).$$

Note this bound does not depend on θ or n.

Lastly,

$$\sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta) - Q_0(\theta)\right| \leq \sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta) - \hat{Q}_n^*(\theta)\right| + \sup_{\theta\in\Theta}\left|\hat{Q}_n^*(\theta) - Q_0(\theta)\right|,$$

where $\sup_{\theta\in\Theta}\left|\hat{Q}_n^*(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0$ by Lemma 13. By $\delta(n)\to 0$, $m(n)\to+\infty$, and

$$\lim_{m\to+\infty} \ln \frac{\underline{f} + \frac{1}{4}(\overline{f}-\underline{f})J\bar{\rho}^m\overline{D}}{\underline{f}} = 0,$$

we have $\sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta) - \hat{Q}_n^*(\theta)\right| \xrightarrow{p} 0$. Thus, $\sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0$. □

Proof of Theorem 5. We are ready to apply Theorem 2.1 in Newey and McFadden, 1994. (1). By the identification assumption 3, $Q_0(\theta)$ is uniquely maximized at θ_0. (2). Θ is compact. (3). As $Prob_j^*(\theta;\tilde{G})$ is also bounded below by $\underline{f}$ and continuous in θ by Lemma 11, $Q_0(\theta)$ is continuous. (4). $\hat{Q}_n(\theta)$ converges in probability to $Q_0(\theta)$, by Lemma 14. Thus, $\hat{\theta}$ is consistent.

To see $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi \xrightarrow{p} G$, note that

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, G) \leq D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G},\hat{\theta}))+D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0))+D(F^{-1}(\hat{G},\theta_0), G).$$

The first term

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G},\hat{\theta})) \to 0 \quad \text{as} \quad m(n)\to\infty.$$

The second term

$$D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0)) \xrightarrow{p} 0, \quad \text{as} \quad \hat{\theta}\xrightarrow{p}\theta_0$$

[18] This bound is not tight.

and F^{-1} is continuous in θ by Lemma 11. The third term

$$D(F^{-1}(\hat{G}, \theta_0), G) \xrightarrow{p} 0, \quad \text{as} \quad \hat{G} \xrightarrow{p} \tilde{G}$$

and F is a homeomorphism by Proposition 5.

□

Proof of Theorem 6

Lemma 15. *If Assumption 2, 3, and 4 hold, then $\hat{\theta}^*$ is asymptotically normal and $\sqrt{n}(\hat{\theta}^* - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$.*

Proof of Lemma 15. We shall first rewrite the estimator as a generalized method of moment estimator. We let $\hat{Prob} = (\hat{Prob}_1, \hat{Prob}_2, \cdots, \hat{Prob}_J)'$ denote the observed frequency of alternatives. Let $\mathbf{1}_p$ denote the cumulative indicator vector that assigns 0 for entries $p_j < p$ and 1 for entries $p_j \geq p$. Estimator $\hat{\theta}^*$ solves the first-order condition of Equation (2.11)

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{g}^*(z_i, \theta, \hat{G}) = 0,$$

where $\hat{G}$ satisfies the moment condition

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{Prob} - y_i) = 0 \tag{2.23}$$

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{G}_j - y_{ij}\mathbf{1}_{p_i}/\hat{Prob}_j) = 0 \quad \text{for all } j \in \mathcal{J}, \tag{2.24}$$

where p_i is the observed selected price for individual i.

For this standard GMM estimator, we can directly invoke Theorem 6.1 in Newey and McFadden, 1994. Note that our $\mathfrak{g}^*$ is their g and our $(\hat{Prob}, \hat{G})$ is their $\hat{\gamma}$ in Newey and McFadden, 1994. Let

$$\mathfrak{m}_1(z_i, \hat{Prob}) = \hat{Prob} - y_i,$$

$$\mathfrak{m}_2(z_i, \hat{Prob}, \hat{G}) = [[\hat{G}_1 - y_{i1}\mathbf{1}_{p_i}/\hat{Prob}_1]', [\hat{G}_2 - y_{i2}\mathbf{1}_{p_i}/\hat{Prob}_2]', \cdots, [\hat{G}_J - y_{J2}\mathbf{1}_{p_i}/\hat{Prob}_J]']'.$$

We stack $\mathfrak{g}^*$, $\mathfrak{m}_1$, $\mathfrak{m}_2$ to form $\tilde{\mathfrak{g}}^*$

$$\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G}) = [\mathfrak{g}^*(z, \theta, \hat{G})', \mathfrak{m}_1(z, \hat{Prob})', \mathfrak{m}_2(z, \hat{Prob}, \hat{G})']'.$$

By the proof of Theorem 5 and Lemma 13, $\hat{\theta}^* \xrightarrow{p} \theta_0$. By the weak law of large numbers, $\hat{G} \xrightarrow{p} \tilde{G}$ and $\hat{Prob} \xrightarrow{p} Prob_0 = Prob^*(\theta_0, \tilde{G})$. By Assumption 4, $\theta_0 \in \Theta^o$. Next, we verify that $\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G})$ is continuously differentiable in $\theta, \hat{Prob}, \hat{G}$.

First, we verify that $\mathfrak{g}^*(z, \theta, \hat{G})$ is continuously differentiable in θ. It suffices to show that $Prob^*(\theta, \hat{G})$ is twice continuously differentiable in θ. As f is twice continuously differentiable in θ, we only need to show that $F^{-1}(\hat{G}, \theta)$ is twice continuously differentiable in θ. By Equation (2.1), (2.2) and f being twice continuously differentiable in θ, $F(G, \theta)$ is twice continuously differentiable in θ and infinitely continuously differentiable in G. Thus, by the implicit function theorem,

$$\nabla_\theta F^{-1}(\tilde{G}, \theta) = -\left[\nabla_G F(G, \theta)\right]^{-1} \nabla_\theta F(G, \theta),$$

where the matrix $\nabla_G F(G, \theta)$ is non-singular by F^{-1} being Lipschitz continuous. Consequently, F^{-1} is twice continuously differentiable in θ.

Next, we verify that $\mathfrak{g}^*(z, \theta, \hat{G})$ is continuously differentiable in $\hat{G}$. It suffices to show that $F^{-1}(\hat{G}, \theta)$ is continuously differentiable in $\hat{G}$. As $F(G, \theta)$ is infinitely continuously differentiable in G and $F^{-1}(\hat{G}, \theta)$ is Lipschitz continuous in $\hat{G}$, we have

$$\nabla_{\hat{G}} F^{-1}(\hat{G}, \theta) = [\nabla_G F(G, \theta)]^{-1}$$

which is continuous in $\hat{G}$. Additionally, $\mathfrak{m}_1$ and $\mathfrak{m}_2$ are infinitely continuously differentiable in all parameters $\theta, \hat{G}, \hat{Prob}$. Consequently, we have show that $\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G})$ is continuously differentiable in $\theta, \hat{Prob}, \hat{G}$.

In addition,

$$\mathbb{E}[\mathfrak{g}^*(z, \theta_0, \tilde{G})] = 0$$

by the first-order condition of Q_0. Since f_j is bounded from 0, $||\mathfrak{g}^*(z, \theta_0, \tilde{G})||$ is finite for each z. Furthermore, as supp(G) is finite, there is only a finite possible values of z. Thus, $\mathbb{E}[||\mathfrak{g}^*(z, \theta_0, \tilde{G})||^2]$ is finite. By $\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G})$ being continuously differentiable in $(\theta, \hat{Prob}, \hat{G})$ and a finite possible values of z,

$$\mathbb{E}[\sup_{\theta, \hat{Prob}, \hat{G}} ||\nabla_{\theta, \hat{Prob}, \hat{G}} \tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G})||] < \infty.$$

The last condition we need is that

$$\mathbb{E}[\nabla_{\theta, \hat{Prob}, \hat{G}} \tilde{\mathfrak{g}}^*(z, \theta_0, Prob_0, \tilde{G})]$$

being nonsingular. The matrix $\nabla_{\theta,\hat{Prob},\hat{G}}\tilde{\mathfrak{g}}^*(z,\theta_0,Prob_0,\tilde{G})$ is

$$\begin{pmatrix} \nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G}) & \mathbf{0} & \nabla_{\hat{G}}\mathfrak{g}^*(z,\theta_0,\tilde{G}) \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \nabla_{\hat{Prob}}\mathfrak{m}_2(z,Prob_0,\tilde{G}) & \boldsymbol{I} \end{pmatrix}.$$

Its expectation being nonsingular is equivalent to $\mathbb{E}\nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G})$ being nonsingular, which is in Assumption 4.

We can write down the variance matrix V by Theorem 6.1 in Newey and McFadden, 1994.

$$\begin{aligned} A(z) =& \mathfrak{g}^*(z,\theta_0,\tilde{G}) + \left(\mathbb{E}\nabla_{\hat{G}}\mathfrak{g}^*(z,\theta_0,\tilde{G})\right)\times \\ & \left[\left(\mathbb{E}\nabla_{\hat{Prob}} m_2(z,Prob_0,\tilde{G})\right)\times \mathfrak{m}_1(z,Prob_0) - \mathfrak{m}_2(z,Prob_0,\tilde{G})\right]. \end{aligned}$$

$$V = \left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G})\right)^{-1} \times \mathbb{E}(A(z)A(z)') \times \left(\left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G})\right)^{-1}\right)'.$$

□

Proof of Theorem 6. Recall $\hat{\theta}$ solves the first-order condition

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{g}(z_i;\hat{\theta},\hat{G},n) = 0.$$

We expand this equation around θ_0 and solve for $\sqrt{n}(\hat{\theta}-\theta_0)$

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \mathfrak{g}(z_i,\bar{\theta},\hat{G},n)\right]^{-1}\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\mathfrak{g}(z_i,\theta_0,\hat{G},n),$$

where the second summation is

$$\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\mathfrak{g}(z_i,\theta_0,\hat{G},n) = \sum_{i=1}^{n}\frac{1}{\sqrt{n}}\left(\mathfrak{g}^*(z_i,\theta_0,\hat{G}) + o_p(\frac{1}{\sqrt{n}})\right) = \sum_{i=1}^{n}\frac{1}{\sqrt{n}}\mathfrak{g}^*(z_i,\theta_0,\hat{G}) + o_p(1)$$

by Assumption 4 (v). Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \mathfrak{g}(z_i,\bar{\theta},\hat{G},n) = \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \mathfrak{g}^*(z_i,\bar{\theta},\hat{G}) + o_p(1).$$

Thus, $\sqrt{n}(\hat{\theta}-\theta_0)$ converges to

$$\left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z;\theta_0,\tilde{G})\right)^{-1}\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\mathfrak{g}^*(z_i,\theta_0,\hat{G}) + o_p(1)$$

which has the same limiting distribution as $\sqrt{n}(\hat{\theta}^*-\theta_0)$. Thus, $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0,V)$ by Lemma 15. □

To see the convergence rate of $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$, note that

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, G) \leq D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G},\hat{\theta}))+D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0))+D(F^{-1}(\hat{G},\theta_0), G).$$

The first term goes to 0 at rate faster than $\sqrt{n}$ by Assumption 4 (v). By the proof of Lemma 15, F^{-1} is continuously differentiable in θ; as Θ is compact, F^{-1} is Lipschitz continuous in θ. As $\hat{\theta} \xrightarrow{p} \theta_0$ at rate $\sqrt{n}$,

$$D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0)) \xrightarrow{p} 0 \quad \text{at rate } \sqrt{n}.$$

The last term converges in probability to 0 at rate $\sqrt{n}$, as $\hat{G} \xrightarrow{p} \tilde{G}$ at rate $\sqrt{n}$ and by Proposition 5.

2.10 Tables

Table 2.3: Simulation Results for Utility Parameters: $N = 5000$

	Functional Contraction			Two-Step Method		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0022	0.0864	0.0864	0.0017	0.0917	0.0916
ξ_2	0.0028	0.0345	0.0345	0.0045	0.0427	0.0429
β	-0.0007	0.0412	0.0411	0.0003	0.0417	0.0416
			DGP 2			
γ	0.0011	0.0858	0.0857	0.0157	0.0982	0.0994
ξ_2	0.0001	0.0344	0.0344	0.0051	0.0510	0.0512
β	-0.0009	0.0427	0.0427	0.0052	0.0434	0.0436
			DGP 3			
γ	-0.0133	0.0707	0.0719	0.1611	0.0984	0.1887
ξ_2	0.0019	0.0317	0.0317	0.0137	0.0363	0.0388
β	-0.0012	0.0414	0.0414	-0.0002	0.0417	0.0417
			DGP 4			
γ	0.0026	0.1544	0.1543	0.0433	0.1666	0.1720
ξ_2	0.0009	0.0306	0.0306	-0.0001	0.0314	0.0314
β	-0.0003	0.0404	0.0404	-0.0002	0.0406	0.0405
			DGP 5			
γ	-0.0054	0.4395	0.4391	-0.0043	0.4274	0.4270
ξ_2	0.0009	0.0304	0.0304	0.0010	0.0305	0.0305
β	0.0002	0.0399	0.0398	0.0003	0.0398	0.0398

Table 2.4: Simulation Results for CDF of log(Price): $N = 5000$

	Func. Contraction		Two-Step Method	
	IBias^2	IMSE	IBias^2	IMSE
		DGP 1		
$F_1(\cdot\|x_{i2}=0)$	0.0002	0.0007	0.0000	0.0040
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0002	0.0000	0.0003
$F_1(\cdot\|x_{i2}=1)$	0.0002	0.0009	0.0000	0.0024
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0000	0.0008
		DGP 2		
$F_1(\cdot\|x_{i2}=0)$	0.0003	0.0009	0.0024	0.0078
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0002	0.0020	0.0023
$F_1(\cdot\|x_{i2}=1)$	0.0004	0.0010	0.0024	0.0056
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0020	0.0029
		DGP 3		
$F_1(\cdot\|x_{i2}=0)$	0.0059	0.0064	0.0209	0.0269
$F_2(\cdot\|x_{i2}=0)$	0.0028	0.0029	0.0493	0.0499
$F_1(\cdot\|x_{i2}=1)$	0.0005	0.0011	0.0037	0.0057
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0143	0.0152
		DGP 4		
$F_1(\cdot\|x_{i2}=0)$	0.0006	0.0011	0.0024	0.0077
$F_2(\cdot\|x_{i2}=0)$	0.0007	0.0008	0.0272	0.0277
$F_1(\cdot\|x_{i2}=1)$	0.0003	0.0012	0.0016	0.0047
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0098	0.0104
		DGP 5		
$F_1(\cdot\|x_{i2}=0)$	0.0014	0.0018	0.0011	0.0053
$F_2(\cdot\|x_{i2}=0)$	0.0014	0.0015	0.0202	0.0206
$F_1(\cdot\|x_{i2}=1)$	0.0007	0.0018	0.0016	0.0055
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0081	0.0084

Note: The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R}\hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

Table 2.5: Simulation Results for Utility Parameters: Removing the Excluded Variable

	$N = 1000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0011	0.2082	0.2080	0.0018	0.0866	0.0865
ξ_2	0.0074	0.0570	0.0574	0.0000	0.0254	0.0253
			DGP 2			
γ	-0.0018	0.2066	0.2064	0.0024	0.1000	0.0999
ξ_2	0.0033	0.0535	0.0535	0.0028	0.0264	0.0265
			DGP 3			
γ	-0.0163	0.1581	0.1587	-0.0043	0.0728	0.0729
ξ_2	0.0061	0.0542	0.0544	0.0007	0.0238	0.0238
			DGP 4			
γ	0.0019	0.3660	0.3656	0.0059	0.1563	0.1563
ξ_2	0.0050	0.0498	0.0500	-0.0005	0.0225	0.0225
			DGP 5			
γ	0.0000	1.0797	1.0786	-0.0146	0.4409	0.4407
ξ_2	0.0014	0.0531	0.0530	-0.0007	0.0233	0.0233

Note: In these specifications, we remove the excluded variable from the selection function, so the parameter β in u_{i1} is not estimated.

Table 2.6: Simulation Results for CDF of log(Price): Removing the Excluded Variable

	$N = 1000$		$N = 5000$	
	IBias^2	IMSE	IBias^2	IMSE
		DGP 1		
$F_1(\cdot\|x_{i2} = 0)$	0.0002	0.0018	0.0002	0.0004
$F_2(\cdot\|x_{i2} = 0)$	0.0001	0.0003	0.0000	0.0001
$F_1(\cdot\|x_{i2} = 1)$	0.0003	0.0019	0.0002	0.0005
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0007	0.0001	0.0002
		DGP 2		
$F_1(\cdot\|x_{i2} = 0)$	0.0004	0.0017	0.0003	0.0006
$F_2(\cdot\|x_{i2} = 0)$	0.0002	0.0004	0.0001	0.0001
$F_1(\cdot\|x_{i2} = 1)$	0.0005	0.0018	0.0003	0.0006
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0007	0.0001	0.0002
		DGP 3		
$F_1(\cdot\|x_{i2} = 0)$	0.0058	0.0073	0.0061	0.0064
$F_2(\cdot\|x_{i2} = 0)$	0.0029	0.0031	0.0028	0.0028
$F_1(\cdot\|x_{i2} = 1)$	0.0006	0.0021	0.0005	0.0008
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0007	0.0000	0.0002
		DGP 4		
$F_1(\cdot\|x_{i2} = 0)$	0.0006	0.0021	0.0006	0.0008
$F_2(\cdot\|x_{i2} = 0)$	0.0008	0.0010	0.0007	0.0007
$F_1(\cdot\|x_{i2} = 1)$	0.0004	0.0024	0.0003	0.0007
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0006	0.0000	0.0002
		DGP 5		
$F_1(\cdot\|x_{i2} = 0)$	0.0014	0.0025	0.0013	0.0016
$F_2(\cdot\|x_{i2} = 0)$	0.0013	0.0015	0.0014	0.0014
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0033	0.0006	0.0012
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0006	0.0001	0.0002

Note: In these specifications, we remove the excluded variable from the selection function. The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

Table 2.7: Simulation Results for Utility Parameters: Misspecifying the Selection Function

	$N = 1000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0793	0.1826	0.1989	-0.0743	0.0806	0.1096
ξ_2	-0.0754	0.0714	0.1038	-0.0752	0.0342	0.0826
β	-0.0309	0.0856	0.0909	-0.0306	0.0392	0.0497
			DGP 2			
γ	-0.0748	0.1864	0.2007	-0.0667	0.0806	0.1045
ξ_2	-0.0714	0.0727	0.1018	-0.0742	0.0340	0.0816
β	-0.0315	0.0902	0.0954	-0.0282	0.0407	0.0495
			DGP 3			
γ	-0.1051	0.1475	0.1810	-0.0940	0.0650	0.1142
ξ_2	-0.0789	0.0698	0.1053	-0.0768	0.0317	0.0830
β	-0.0323	0.0887	0.0943	-0.0293	0.0398	0.0494
			DGP 4			
γ	-0.0746	0.3249	0.3330	-0.0584	0.1442	0.1554
ξ_2	-0.0776	0.0679	0.1030	-0.0755	0.0307	0.0814
β	-0.0263	0.0900	0.0937	-0.0222	0.0392	0.0450
			DGP 5			
γ	0.0029	0.9169	0.9160	-0.0606	0.4177	0.4217
ξ_2	-0.0827	0.0667	0.1063	-0.0801	0.0302	0.0856
β	-0.0315	0.0847	0.0902	-0.0266	0.0381	0.0465

Note: In these specifications, we misspecify the selection model, assuming that the error term ε_i is drawn from $Logistic(0, 1)$.

Table 2.8: Simulation Results for CDF of log(Price): Misspecifying the Selection Function

	$N = 1000$		$N = 5000$	
	IBias^2	IMSE	IBias^2	IMSE
	DGP 1			
$F_1(\cdot\|x_{i2} = 0)$	0.0004	0.0029	0.0002	0.0007
$F_2(\cdot\|x_{i2} = 0)$	0.0001	0.0006	0.0001	0.0002
$F_1(\cdot\|x_{i2} = 1)$	0.0004	0.0032	0.0002	0.0009
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0013	0.0001	0.0003
	DGP 2			
$F_1(\cdot\|x_{i2} = 0)$	0.0006	0.0033	0.0005	0.0010
$F_2(\cdot\|x_{i2} = 0)$	0.0002	0.0006	0.0001	0.0002
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0037	0.0004	0.0010
$F_2(\cdot\|x_{i2} = 1)$	0.0003	0.0015	0.0001	0.0004
	DGP 3			
$F_1(\cdot\|x_{i2} = 0)$	0.0062	0.0087	0.0061	0.0066
$F_2(\cdot\|x_{i2} = 0)$	0.0028	0.0032	0.0028	0.0029
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0033	0.0005	0.0011
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0013	0.0001	0.0003
	DGP 4			
$F_1(\cdot\|x_{i2} = 0)$	0.0008	0.0034	0.0006	0.0012
$F_2(\cdot\|x_{i2} = 0)$	0.0008	0.0012	0.0007	0.0008
$F_1(\cdot\|x_{i2} = 1)$	0.0006	0.0046	0.0003	0.0012
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0011	0.0001	0.0003
	DGP 5			
$F_1(\cdot\|x_{i2} = 0)$	0.0014	0.0034	0.0014	0.0019
$F_2(\cdot\|x_{i2} = 0)$	0.0014	0.0018	0.0014	0.0015
$F_1(\cdot\|x_{i2} = 1)$	0.0008	0.0058	0.0007	0.0018
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0011	0.0001	0.0003

Note: In these specifications, we misspecify the selection model, assuming that the error term ε_i is drawn from $Logistic(0, 1)$. The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

BIBLIOGRAPHY

Ahn, Hyungtaik and James L Powell (1993). "Semiparametric estimation of censored selection models with a nonparametric selection mechanism". In: *Journal of Econometrics* 58.1-2, pp. 3–29.

Allen, Jason, Robert Clark, Brent Hickman, et al. (2024). "Resolving failed banks: Uncertainty, multiple bidding and auction design". In: *Review of Economic Studies* 91.3, pp. 1201–1242.

Allen, Jason, Robert Clark, and Jean-François Houde (2014). "Price dispersion in mortgage markets". In: *The Journal of Industrial Economics* 62.3, pp. 377–416.

– (2019). "Search frictions and market power in negotiated-price markets". In: *Journal of Political Economy* 127.4, pp. 1550–1598.

Andrews, Donald WK and Marcia MA Schafgans (1998). "Semiparametric estimation of the intercept of a sample selection model". In: *The Review of Economic Studies* 65.3, pp. 497–517.

Arellano, Manuel and Stéphane Bonhomme (2017). "Quantile selection models with an application to understanding changes in wage inequality". In: *Econometrica* 85.1, pp. 1–28.

Asker, John and Estelle Cantillon (2008). "Properties of scoring auctions". In: *The RAND Journal of Economics* 39.1, pp. 69–85.

Athey, Susan and Philip A Haile (2002). "Identification of standard auction models". In: *Econometrica* 70.6, pp. 2107–2140.

– (2007). "Nonparametric approaches to auctions". In: *Handbook of econometrics* 6, pp. 3847–3965.

Baricz, Árpád (2008). "Mills' ratio: Monotonicity patterns and functional inequalities". In: *Journal of Mathematical Analysis and Applications* 340.2, pp. 1362–1370.

Bayer, Patrick, Shakeeb Khan, and Christopher Timmins (2011). "Nonparametric identification and estimation in a roy model with common nonpecuniary returns". In: *Journal of Business & Economic Statistics* 29.2, pp. 201–215.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995). "Automobile prices in market equilibrium". In: *Econometrica* 63.4, pp. 841–890. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2171802 (visited on 01/20/2024).

Berry, Steven T (1994). "Estimating discrete-choice models of product differentiation". In: *The RAND Journal of Economics*, pp. 242–262.

Borjas, GJ (1987). "Self-selection and the Earnings of Immigrants". In: *American Economic Review* 77, pp. 531–553.

Buchholz, Nicholas et al. (2020). “The value of time: Evidence from auctioned cab rides”. In: *CEPR Discussion Paper No. DP14666*.

Campo, Sandra (2012). “Risk aversion and asymmetry in procurement auctions: Identification, estimation and application to construction procurements”. In: *Journal of Econometrics* 168.1, pp. 96–107.

Canay, Ivan A, Magne Mogstad, and Jack Mountjoy (2024). “On the use of outcome tests for detecting bias in decision making”. In: *Review of Economic Studies* 91.4, pp. 2135–2167.

Chen, Songnian and Shakeeb Khan (2003). “Semiparametric estimation of a heteroskedastic sample selection model”. In: *Econometric Theory* 19.6, pp. 1040–1064.

Chernozhukov, Victor, Iván Fernández-Val, and Siyi Luo (2023). *Distribution regression with sample selection and UK wage decomposition*. Tech. rep. cemmap working paper.

Cicala, Steve (2015). “When does regulation distort costs? lessons from fuel procurement in us electricity generation”. In: *American Economic Review* 105.1, pp. 411–444.

Compiani, Giovanni and Yuichi Kitamura (2016). “Using mixtures in econometric models: a brief review and some new results”. In: *The Econometrics Journal* 19.3, pp. C95–C127.

Cosconati, Marco et al. (2024). “Competing under Information Heterogeneity: Evidence from auto insurance”. In: *Working paper*.

Crawford, Gregory S, Nicola Pavanini, and Fabiano Schivardi (2018). “Asymmetric information and imperfect competition in lending markets”. In: *American Economic Review* 108.7, pp. 1659–1701.

D’Haultfœuille, Xavier, Arnaud Maurel, and Yichong Zhang (2018). “Extremal quantile regressions for selection models and the black–white wage gap”. In: *Journal of Econometrics* 203.1, pp. 129–142.

d’Haultfoeuille, Xavier and Arnaud Maurel (2013a). “Another look at the identification at infinity of sample selection models”. In: *Econometric Theory* 29.1, pp. 213–224.

– (2013b). “Inference on an extended Roy model, with an application to schooling decisions in France”. In: *Journal of Econometrics* 174.2, pp. 95–106.

Das, Mitali, Whitney K Newey, and Francis Vella (2003). “Nonparametric estimation of sample selection models”. In: *The Review of Economic Studies* 70.1, pp. 33–58.

De Silva, Dakshina G, Georgia Kosmopoulou, and Carlos Lamarche (2009). “The effect of information on the bidding and survival of entrants in procurement auctions”. In: *Journal of Public Economics* 93.1-2, pp. 56–72.

Dubé, Jean-Pierre and Sanjog Misra (2023). "Personalized pricing and consumer welfare". In: *Journal of Political Economy* 131.1, pp. 131–189.

Fernández-Val, Ivan, Aico van Vuuren, and Francis Vella (2024). "Nonseparable sample selection models with censored selection rules". In: *Journal of Econometrics* 240.2, p. 105088. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2021.01.009. URL: https://www.sciencedirect.com/science/article/pii/S0304407621000567.

Flambard, Véronique and Isabelle Perrigne (2006). "Asymmetry in procurement auctions: Evidence from snow removal contracts". In: *The Economic Journal* 116.514, pp. 1014–1036.

French, Eric and Christopher Taber (2011). "Identification of models of the labor market". In: *Handbook of labor economics*. Vol. 4. Elsevier, pp. 537–617.

Goldberg, Pinelopi Koujianou (1996). "Dealer price discrimination in new car purchases: Evidence from the consumer expenditure survey". In: *Journal of Political Economy* 104.3, pp. 622–654.

Gronau, Reuben (1974). "Wage comparisons–A selectivity bias". In: *Journal of political Economy* 82.6, pp. 1119–1143.

Guerre, Emmanuel and Yao Luo (2019). "Nonparametric identification of first-price auction with unobserved competition: A density discontinuity framework". In: *arXiv preprint arXiv:1908.05476*.

Heckman, James J (1974). "Shadow prices, market wages, and labor supply". In: *Econometrica: journal of the econometric society*, pp. 679–694.

– (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". In: *Annals of economic and social measurement, volume 5, number 4*. NBER, pp. 475–492.

– (1979). "Sample selection bias as a specification error". In: *Econometrica* 47, pp. 153–161.

Heckman, James J and Bo E Honore (1990). "The empirical content of the Roy model". In: *Econometrica: Journal of the Econometric Society*, pp. 1121–1149.

Heckman, James J and Guilherme Sedlacek (1985). "Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market". In: *Journal of political Economy* 93.6, pp. 1077–1125.

Heckman, James J and Edward J Vytlacil (2007). "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". In: *Handbook of econometrics* 6, pp. 4779–4874.

Hendricks, Kenneth and Robert H Porter (1988). "An empirical study of an auction with asymmetric information". In: *The American Economic Review*, pp. 865–883.

Hortaçsu, Ali et al. (2019). "Does strategic ability affect efficiency? Evidence from electricity markets". In: *American Economic Review* 109.12, pp. 4302–4342.

Hu, Yingyao (2017). "The Econometrics of Unobservables–Latent Variable and Measurement Error Models and Their Applications in Empirical Industrial Organization and Labor Economics". In: *Manuscript in preparation*.

Ichimura, Hidehiko (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". In: *Journal of econometrics* 58.1-2, pp. 71–120.

Klein, Roger W and Richard H Spady (1993). "An efficient semiparametric estimator for binary response models". In: *Econometrica: Journal of the Econometric Society*, pp. 387–421.

Komarova, Tatiana (2013). "A new approach to identifying generalized competing risks models with application to second-price auctions". In: *Quantitative Economics* 4.2, pp. 269–328.

Krasnokutskaya, Elena, Kyungchul Song, and Xun Tang (2020). "The role of quality in internet service markets". In: *Journal of Political Economy* 128.1, pp. 75–117.

Lee, Ji Hyung and Byoung G Park (2023). "Nonparametric identification and estimation of the extended Roy model". In: *Journal of Econometrics* 235.2, pp. 1087–1113.

Lee, Lung-Fei (1978). "Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables". In: *International economic review*, pp. 415–433.

– (1982). "Some approaches to the correction of selectivity bias". In: *The Review of Economic Studies* 49.3, pp. 355–372.

– (1983). "Generalized econometric models with selectivity". In: *Econometrica: Journal of the Econometric Society*, pp. 507–512.

Lewis, Gregory and Patrick Bajari (2011). "Procurement contracting with time incentives: Theory and evidence". In: *The Quarterly Journal of Economics* 126.3, pp. 1173–1211.

McKelvey, Richard D and Thomas R Palfrey (1995). "Quantal response equilibria for normal form games". In: *Games and Economic Behavior* 10.1, pp. 6–38.

Meilijson, Isaac (1981). "Estimation of the lifetime distribution of the parts from the autopsy statistics of the machine". In: *Journal of Applied Probability* 18.4, pp. 829–838.

Mourifie, Ismael, Marc Henry, and Romuald Meango (2020). "Sharp bounds and testability of a Roy model of STEM major choices". In: *Journal of Political Economy* 128.8, pp. 3220–3283.

Nakabayashi, Jun (2013). "Small business set-asides in procurement auctions: An empirical analysis". In: *Journal of Public Economics* 100, pp. 28–44.

Newey, Whitney K (2007). “Nonparametric continuous/discrete choice models”. In: *International Economic Review* 48.4, pp. 1429–1439.

– (2009). “Two-step series estimation of sample selection models”. In: *The Econometrics Journal* 12.suppl_1, S217–S229.

Newey, Whitney K and Daniel McFadden (1994). “Large sample estimation and hypothesis testing”. In: *Handbook of Econometrics* 4, pp. 2111–2245.

Roy, Andrew Donald (1951). “Some thoughts on the distribution of earnings”. In: *Oxford Economic Papers* 3.2, pp. 135–146.

Sagl, Stephan (2023). “Dispersion, discrimination, and the price of your pickup”. In: *Working paper*.

Salz, Tobias (2022). “Intermediation and competition in search markets: An empirical case study”. In: *Journal of Political Economy* 130.2, pp. 310–345.

Takahashi, Hidenori (2018). “Strategic design under uncertain evaluations: Structural analysis of design-build auctions”. In: *The RAND Journal of Economics* 49.3, pp. 594–618.

Thompson, Anthony C (1963). “On certain contraction mappings in a partially ordered vector space.” In: *Proceedings of the American Mathematical Society* 14.3, pp. 438–443.

Vella, Francis (1998). “Estimating models with sample selection bias: a survey”. In: *Journal of Human Resources*, pp. 127–169.

Willis, Robert J and Sherwin Rosen (1979). “Education and self-selection”. In: *Journal of Political Economy* 87.5, Part 2, S7–S36.

Yoganarasimhan, Hema (2016). “Estimation of beauty contest auctions”. In: *Marketing Science* 35.1, pp. 27–54.

Chapter 3

IMPLEMENTING RANDOMIZED ALLOCATION RULES WITH OUTCOME-CONTINGENT TRANSFERS

Liu, Yi and Fan Wu (2024). "Implementing randomized allocation rules with outcome-contingent transfers". In: *Journal of Economic Theory* 220, p. 105878. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2024.105878. URL: https://www.sciencedirect.com/science/article/pii/S002205312400084X.

3.1 Introduction

A central topic in accounting is earnings management. Each year, firms are required to report their annual economic activities to an auditing company in order to compile financial statements. Auditors exercise a certain level of discretion in this task due to reputation and potential legal implications. These financial statements hold significant importance for the firm, as they directly impact future financing costs. Furthermore, taxes are calculated based on these statements. In practice, some firms resort to tactics such as window dressing, which involves inflating their reports to present a more favorable financial picture. Firms may also engage in earning manipulation, deliberately under reporting their performance to minimize tax liabilities. This prompts a natural question: is there a tax regime that could encourage firms to report truthfully? If such a regime exists, what does it look like?

Similar challenges arise in the field of political science. For instance, on an annual basis, each province in China reports its economic growth, fiscal surplus, expected annual budget, and other relevant information to the Bureau of Statistics. This report comprises high-dimensional data, encompassing all aspects of economic activity. The Bureau of Statistics evaluates the overall economic condition of each province. The evaluation is of particular interest to the provinces as it may influence their economic policies in the future.[1] Based on the evaluation, the central government

[1]For example, many of China's economic special zones and new areas are selected due to fast economic growth, including Shenzhen Special Economic Zone, Shanghai Pudong New Area, and Zhuhai Hengqin New Area in Guangdong. Once established as special economic zones or new areas, a district enjoys special policy treatment, including tax reduction, relaxation of market access, simplification of administrative approval, etc.

determines the fiscal transfers between provinces.[2] Is it possible for the central government to devise a transfer scheme that incentivizes every province to truthfully report its economic condition?

We analyze these problems using a mechanism design approach. We study a model where an agent knows the underlying state of the world, θ, which belongs to a set Θ. The agent reports a state to an exogenous *allocation rule*, a function $\pi\colon \Theta \to \Delta(X)$ mapping each state to a distribution over a finite outcome space X. The agent's valuation of the outcome x in state θ is given by $v(\theta, x)$. A planner designs an outcome-contingent transfer $t\colon X \to \mathbb{R}$ describing the agent's monetary payoff as a function of the outcome. Our research question is whether there exists an outcome-contingent transfer to induce the agent to report the state truthfully.

In the earnings management example, the agent is a firm that reports its financial situation θ to an auditing company. The company's auditing produces a financial statement x. The allocation rule summarizes the auditor's practices and protocol. Then the government collects tax $t(x)$ based on the financial statement. In the fiscal transfer example, the agent is a province that reports its economic activities θ to the Bureau of Statistics. The Bureau of Statistics assigns an economic evaluation x to the province. The allocation rule summarizes the evaluation procedure. (The randomness in the evaluation rule is to reduce a province's incentive to manipulate its report.) Then the central government assigns a transfer $t(x)$ based on the evaluation.[3]

In these two examples, the transfer's contingency on the evaluation (financial statement) stems from the fact that different entities are responsible for evaluation (auditing) and transfer (tax) assignments. Moreover, as the provincial (firm's) economic activity is high-dimensional, part of the Bureau of Statistics' (auditor's) job is to simplify the task of transfer (tax) assignment for the central government.

The key novelty in our model is that the allocation rule allows for randomization and that transfers depend on the allocation outcome rather than directly on the report. When an allocation rule is deterministic, whether the transfer depends on the report or the outcome is irrelevant. This is also known as the taxation principle or tariff principle. In real life, we observe tariffs more often than direct revelation mechanisms, because the set of possible type spaces may be hard to describe in

[2]The magnitude of transfer for each province is roughly ten billion dollars. The aggregate transfer is roughly a trillion dollars.

[3]The transfer's contingency on evaluations and the randomness in the evaluation/allocation rule also arise from confidentiality concerns. If the transfer were to depend directly on the state, it would reveal too much information about the state, as the transfer is publicly observable.

reality (Tadelis and Segal, 2005). So simplicity is a prominent advantage of tariffs and our mechanism directly inherits this advantage when the type space is large.

Another motivation for our mechanism is that it could be useful in other mechanism design problems. The standard mechanism with report-contingent transfer specifies a mapping that assigns to each type an allocation and a transfer. Our model, instead, decouples this mapping into two functions, the allocation rule $\pi\colon \Theta \to \Delta(X)$ and the transfer rule $t\colon X \to \mathbb{R}$. This decomposition also appears in the canonical mechanism in Doval and Skreta, 2022. In their leading example (see their Section 3.1), the canonical mechanism is a mechanism with outcome-contingent transfers.[4] Our paper answers the question of when such transfers exist under the truthful report (IC) constraint.

The taxation principle states that when an allocation rule is deterministic, whether the transfer depends on the report or the outcome is irrelevant. However, if the allocation rule is randomized, we show that it is harder to implement with outcome-contingent transfers (see Observation 1). We say that the allocation rule is *implementable* with outcome-contingent transfers if there exist outcome-contingent transfers such that it is optimal for the agent to report truthfully in each state.

Our main result is a characterization of implementable allocation rules. We collect the agent's valuation of all outcomes in state θ into a $|X|$-dimensional vector $v(\theta)$. For each pair of states (θ, θ') we define the *allocation difference* as the difference in probabilities

$$d\pi(\theta, \theta') = \pi(\theta) - \pi(\theta')$$

and the *valuation loss* as the agent's difference in valuation in the two states:

$$\mathrm{vl}(\theta, \theta') = d\pi(\theta, \theta') \cdot v(\theta).$$

We collect the allocation differences into a $|X| \times |\Theta|^2$ dimensional matrix $D\pi$ consisting columns of $d\pi(\theta, \theta')$, and all valuation losses $\mathrm{vl}(\theta, \theta')$ into a valuation loss vector VL, by ordering pairs of states (θ, θ') in the same order.

Recall that a matrix's positive kernel $\ker_+$ is the intersection of the kernel and the positive orthant. We show that the allocation rule is implementable if and only if VL lies in the dual cone of $\ker_+(D\pi)$ (Theorem 7). Moreover, we offer another

[4] In their leading example, the disclosure rule maps a report to a distribution over posterior beliefs, and the price depends only on the realized posterior belief. Their posterior belief is our allocation outcome and their price is our transfer.

geometric characterization. Each pair of distinct states (θ, θ') determines an allocation difference $d\pi(\theta, \theta')$ and a valuation loss $\text{vl}(\theta, \theta')$. This $(d\pi(\theta, \theta'), \text{vl}(\theta, \theta'))$ corresponds to a point in $\mathbb{R}^{|X|} \times \mathbb{R}$. For all such points, we can construct their convex envelope $conv\colon \mathbb{R}^{|X|} \to \mathbb{R}$. We show that the allocation rule is implementable if and only if the convex envelope's intercept $conv(\mathbf{0})$, the value of $conv$ evaluated at $\mathbf{0}$, is weakly positive (Theorem 7).

Moreover, when the allocation rule is implementable with outcome-contingent transfers, we show how transfer payments can be constructed (Proposition 6). In particular, we show that transfers can be recovered from the subgradient of the convex envelope at $d\pi = \mathbf{0}$. When the convex envelope intercept is strictly positive, we show that the allocation rule is strictly implementable (Proposition 7).[5]

In addition, we show that the classic cyclic monotonicity condition of Rochet, 1987 is a necessary condition for implementation in our setup as well (Observation 1). However, without further assumptions on the valuation structure and the allocation rule, cyclic monotonicity is not sufficient in general. Yet we show that cyclic monotonicity is also sufficient when the allocation measures $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly independent (Proposition 8). Additionally, when there are fewer than exactly four states, cyclic monotonicity is also sufficient (Proposition 10). Furthermore, when $\{\pi(\theta)\}_{\theta\in\Theta}$ are convex dependent, we show that it is without loss to only check whether one candidate transfer can implement the allocation rule (Proposition 9).

Literature Review

Our paper contributes to the literature on implementation by studying randomized allocation rules with outcome-contingent transfers. The implementation literature has explored when allocation rules can be truthfully implemented by transfer that depends on the *report*; see Roberts (1979), Rochet (1987), McAfee and McMillan (1988), Jehiel, Moldovanu, and Stacchetti (1999), Gui, Müller, and Vohra (2004), Saks and Yu (2005), Bikhchandani et al. (2006), Müller, Perea, and Wolf (2007), Archer and Kleinberg (2008), Ashlagi et al. (2010), Bergemann, Morris, and Takahashi (2012), Carroll (2012), Carbajal and Müller (2015), Carbajal and Müller (2017), Kushnir and Lokutsievskiy (2021), and Frongillo and Kash (2021).

For single agent settings, Myerson (1981) shows the implementability condition is the subgradient condition in a one-dimensional continuous-type environment.

[5] We say that the allocation rule is *strictly implementable* with outcome-contingent transfers if there exist outcome-contingent transfers such that it is strictly optimal for the agent to report truthfully in each state.

Müller, Perea, and Wolf (2007) and Archer and Kleinberg (2008) propose several equivalent conditions. Rochet (1987) studies when an allocation rule is implementable in dominant strategy mechanisms. He shows that the cyclic monotonicity condition is sufficient and necessary for an allocation rule to be implementable. Bergemann, Morris, and Takahashi (2012) analyze this implementability problem in Bayesian incentive-compatible mechanisms.

In quasilinear environments with a complete domain, Roberts (1979) shows that a positive association of differences is necessary and sufficient for dominant-strategy incentive compatibility. In addition, he derives another characterization in terms of affine maximizers. For a selection of restricted domains, Bikhchandani et al. (2006) characterize dominant-strategy incentive compatibility by weak (cyclic) monotonicity. Gui, Müller, and Vohra (2004) notice that this result holds for the unrestricted domain and for every cube. Saks and Yu (2005) extend this result to any convex multi-dimensional type space. Ashlagi et al. (2010) shows that if the closure of a domain is not convex, then there exists a finite-valued monotone allocation rule that is not implementable. Several more recent works (Kushnir and Lokutsievskiy, 2021; Carbajal and Müller, 2015; Carbajal and Müller, 2017) also identify some conditions under which weak monotonicity (2-cycle monotonicity) is sufficient to implement the allocation rule. Frongillo and Kash (2021) provide a unified framework nesting mechanisms and scoring rules and characterize scoring rules for non-convex sets of distributions.

From a modeling perspective, our work is closely related to the literature of Bayesian persuasion. Our randomized allocation rule can equivalently be seen as a Blackwell experiment.[6] Perez-Richet and Skreta (2022) study the receiver-optimal Blackwell experiment when the sender can falsify the state of the world as the input of the experiment at some cost. Lin and Liu (2024) study when a Blackwell experiment is credible, i.e., when the sender cannot profitably deviate to another experiment while fixing the marginal distribution of realizations. Their credibility also boils down to a cyclic monotonicity condition. Yet their cyclic monotonicity condition is ex-post rather than ex-ante.

[6]Nguyen and Tan (2021) study a model of Bayesian persuasion where the sender does not observe the underlying state, commits to a Blackwell experiment, and privately observes the experiment realization. The sender can misrepresent the experiment's realization with some lying cost. The cost depends on both the experiment realization and the message sent.

3.2 Model

Primitives. We are given a finite set Θ of states. (Our main result Theorem 7 has two sufficient and necessary conditions. The cone condition only applies to finite state spaces. For infinite state space, the envelope condition still holds.) An agent observes the state and sends a report to a predetermined allocation rule. An *allocation rule* π maps a reported state to a distribution over a finite set of outcomes $X = \{x_1, \ldots, x_n\}$, i.e., $\pi \colon \Theta \to \Delta(X)$. Thus each $\pi(\theta)$ is a probability measure over the outcome space X. We denote by $\pi(x|\theta)$ the probability assigned to x by the measure $\pi(\theta)$.

$$\pi(\theta) = (\pi(x_1|\theta), \pi(x_2|\theta), \cdots, \pi(x_n|\theta))^\top.$$

The allocation rule can be seen as outside of the planner's influence, as in our motivating examples, or can be interpreted as the planner's objective. The agent's valuation for outcome x in state θ is equal to $v(\theta, x)$. We write $v(\theta)$ as the $|X|$-dimensional vector consisting of entries $v(\theta, x)$ for all $x \in X$,

$$v(\theta) = (v(\theta, x_1), v(\theta, x_2), \cdots, v(\theta, x_n))^\top.$$

Transfer Design. The agent's total payoff is linear in the valuation and the transfer. The transfer depends only on the outcome and we let $t(x)$ denote the transfer to the agent given outcome x. We use T to denote the n-dimensional transfer vector consisting of entries $t(x)$ for all $x \in X$,

$$T = (t(x_1), t(x_2), \cdots, t(x_n))^\top.$$

We say that the allocation rule is *implementable with outcome-contingent transfers* if there exists a vector T such that for all $\theta,\ \theta' \in \Theta$,

$$\pi(\theta) \cdot (v(\theta) + T) \geq \pi(\theta') \cdot (v(\theta) + T).$$

Compared to the standard implementation with report-contingent transfers, our notion has the additional requirement that the transfer is linear in the allocation rule π. That is, the report-contingent transfer takes the form of $\pi(\theta) \cdot T$.

For any transfer vector T that satisfies this incentive compatibility constraint, any translation $T + (c, c, \cdots, c)$ also satisfies the condition. Thus, we are free to impose an ex-ante budget balance condition. For example, suppose we are given a prior distribution over the states. Given truthful reports, the allocation rule induces a distribution over outcomes. Then we can impose an ex-ante budget balance

condition, i.e., $E_x t(x) = 0$ where E_x is the expectation with respect to the random outcome x.

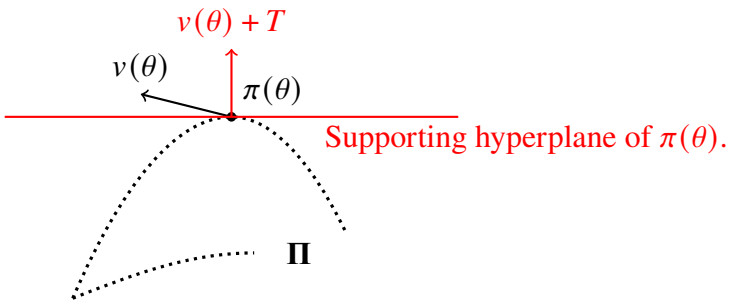


Figure 3.1: The Supporting Hyperplane

Problem Reformulation. Next, we reformulate the implementation problem into the following geometric form. Given the set $\mathbf{\Pi} = \{\pi(\theta)|\theta \in \Theta\}$ in $\mathbb{R}^n$, we associate to each vector $\pi(\theta)$ the vector $v(\theta) \in \mathbb{R}^n$. We ask if there exists a vector $T \in \mathbb{R}^n$ such that for all $\pi(\theta)$, $T + v(\theta)$ is the outer normal of a supporting hyperplane of the set $\mathbf{\Pi}$ at $\pi(\theta)$. That is, for all θ, $\theta' \in \Theta$,

$$[\pi(\theta) - \pi(\theta')] \cdot (v(\theta) + T) \geq 0.$$

Informally, we want a common adjustment vector T such that for all the points in $\mathbf{\Pi}$, the new vector $v + T$ is the outer normal of a supporting hyperplane of the set $\mathbf{\Pi}$ (see Figure 3.1).

3.3 Main Results

We first provide an example where the allocation rule is implementable with standard report-contingent transfers but not with outcome-contingent ones.

Example 2. There are four possible types and two outcomes: $\Theta = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, $X = \{x_1, x_2\}$. The allocation rule is

$$\pi(\theta) = (\theta, 1 - \theta)^\top.$$

The valuation vector is

$$v(\theta) = (\theta, 0)^\top.$$

Note that report-contingent transfer $\tilde{t}(\theta) = -\frac{1}{2}\theta^2$ can implement the allocation rule.

We show that π is not implementable with outcome-contingent transfers and the agent always has an incentive to misreport. For the agent with type θ, the payoff

difference between truthful reporting and misreporting θ' is

$$(\theta - \theta')(\theta + t(x_1) - t(x_2)).$$

If $t(x_1)-t(x_2)+\theta > 0$, the agent would prefer to report $\theta' = 1$. If $t(x_1)-t(x_2)+\theta < 0$, the agent would prefer to report $\theta' = 0$. The agent with type θ that is not 0 or 1 would truthfully report if and only if $\theta = t(x_2) - t(x_1)$. Thus, any transfer that elicits type $1/3$ cannot elicit type $2/3$. This example shows that our implementation notion is stronger than the standard one.

We call

$$d\pi(\theta, \theta') = \pi(\theta) - \pi(\theta')$$

the *allocation difference* between state θ and θ'. We define the *valuation loss* for state θ when inputting θ' to be the agent's expected loss on the valuation

$$\mathrm{vl}(\theta, \theta') = d\pi(\theta, \theta') \cdot v(\theta).$$

We can calculate the agent's expected deviation loss in state θ when reporting θ' as $\mathrm{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T$. Then the incentive compatibility constraint can be written as for all $\theta,\ \theta' \in \Theta$,

$$\mathrm{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T \geq 0.$$

Let $\mathrm{VL} \in R^{|\Theta|\times|\Theta|}$ denote the vector consisting entries $\mathrm{vl}(\theta, \theta')$ by numerating all (θ, θ') pairs. We let $D\pi$ denote a $|X| \times |\Theta|^2$ dimensional matrix consisting columns $d\pi(\theta, \theta')$ with (θ, θ') arranged in the same order as VL. We define the *positive kernel* of $D\pi$ to be

$$\ker_+(D\pi) = \left\{ \lambda \in \mathbb{R}_+^{|\Theta|\times|\Theta|} : \sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} d\pi(\theta, \theta') = 0 \right\}.$$

This set is non-empty as we can set $\lambda_{\theta_1\theta_2} = \lambda_{\theta_2\theta_1} = 1$ and all other entries to be zero. Since the positive kernel is the intersection between the kernel of $D\pi$ (a linear subspace) and the nonnegative orthant, it is a finitely generated convex cone. Its dual cone is given by

$$[\ker_+(D\pi)]^* = \{y \in \mathbb{R}^{|\Theta|\times|\Theta|} \,|\, y \cdot \lambda \geq 0, \forall \lambda \in \ker_+(D\pi)\}.$$

We append the valuation loss to the allocation difference vector to get a new set of vectors

$$D = \{(d\pi(\theta, \theta'), \mathrm{vl}(\theta, \theta')) | \theta \neq \theta' \in \Theta\},$$

which we call the *difference set*. We let $conv(D)$ denote the *convex envelope* of the set D

$$conv(D)(\cdot) = \sup\{g(\cdot)|g\colon \mathbb{R}^n \to \mathbb{R}\ is\ convex\ and\ g(d\pi(\theta,\theta')) \leq \mathrm{vl}(\theta,\theta'), \forall \theta \neq \theta' \in \Theta\}.$$

We call $conv(D)(\mathbf{0})$ the *convex envelope intercept*.

Now we are ready to characterize the implementation condition.

Theorem 7. *The following are equivalent.*

1. *The allocation rule is implementable with outcome-contingent transfers.*
2. $\mathrm{VL} \in [\ker_+(D\pi)]^*$.
3. *The convex envelope intercept is weakly positive.*

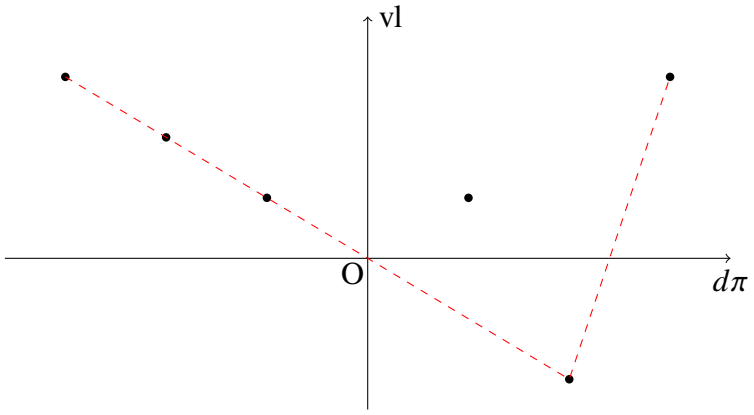


Figure 3.2: An Illustration of Convex Envelope

We provide a geometric example to use the last condition in Theorem 7 to check implementation. Suppose there are three states $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and two outcomes. The allocation rule is $\pi(\theta_1) = (1,0)^T, \pi(\theta_2) = (\frac{2}{3}, \frac{1}{3})^T$ and $\pi(\theta_3) = (0,1)^T$. The agent's valuations are $v(\theta_1) = (1,0)^T, v(\theta_2) = (1,2)^T, v(\theta_3) = (0,1)^T$. Then we construct the difference set

$$D = \{((\frac{1}{3}, -\frac{1}{3}), \frac{1}{3}), ((-\frac{1}{3}, \frac{1}{3}), \frac{1}{3}), ((1,-1),1), ((-1,1),1), ((\frac{2}{3}, -\frac{2}{3}), -\frac{2}{3}), ((-\frac{2}{3}, \frac{2}{3}), \frac{2}{3})\}.$$

Since the entries of $d\pi(\theta,\theta')$ sum up to zero, we can identify each $d\pi$ with an element in $\mathbb{R}$ and plot them in Figure 3.2. The red dashed line is the convex envelope of D and the value of the convex envelope evaluated at $\mathbf{0}$ is 0. So by Theorem 7, the allocation rule is implementable with outcome-contingent transfers.

We illustrate the intuition for the necessity of the convex envelope intercept condition. Suppose $\Theta = \{\theta_1, \theta_2, \cdots\}$ and $\text{vl}(\theta_1, \theta_2) + \text{vl}(\theta_2, \theta_1) < 0$ (see the left panel of Figure 3.3). Then, $\text{vl}(\theta_1, \theta_2) + \text{vl}(\theta_2, \theta_1) < 0$ implies $conv(D)(\mathbf{0}) < 0$, as $d\pi(\theta_1, \theta_2) = -d\pi(\theta_2, \theta_1)$. Similar to Example 2, any transfer T that elicits θ_1 cannot elicit θ_2. To elicit θ_1, we must have $d\pi(\theta_1, \theta_2) \cdot T + \text{vl}(\theta_1, \theta_2) \geq 0$. As $d\pi(\theta_1, \theta_2) = -d\pi(\theta_2, \theta_1)$, the transfer T must have the opposite effect on type θ_2. That is, whenever we use some transfer to ensure $d\pi(\theta_1, \theta_2) \cdot T + \text{vl}(\theta_1, \theta_2) \geq 0$, this leads to $d\pi(\theta_2, \theta_1) \cdot T + \text{vl}(\theta_2, \theta_1) < 0$, as the intercept is preserved (see the right panel of Figure 3.3). In fact, the condition $\text{vl}(\theta_1, \theta_2) + \text{vl}(\theta_2, \theta_1) \geq 0$ is exactly the weak monotonicity and so it must hold for implementation. But the opposing effect around the intercept is driving the necessity of $conv(D)(\mathbf{0}) \geq 0$. The intuition carries over in general such that if $conv(D)(\mathbf{0}) < 0$, for all transfer T, at least one pair (θ_i, θ_j) has $d\pi(\theta_i, \theta_j) \cdot T + \text{vl}(\theta_i, \theta_j) < 0$.

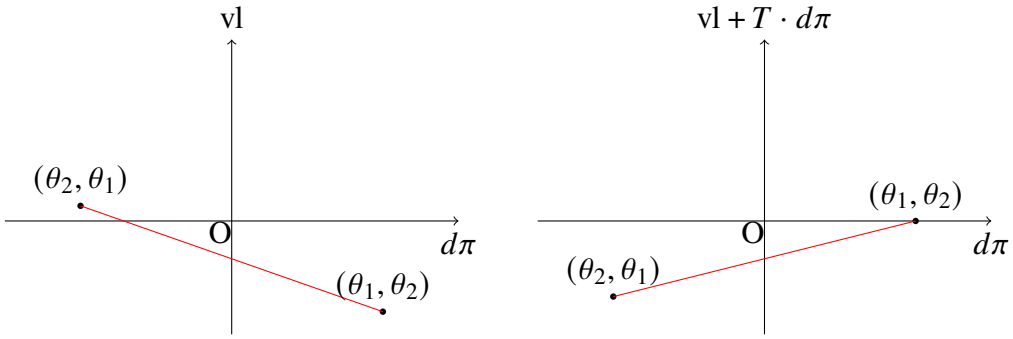


Figure 3.3: The Effect of Transfer

Next, we prove the necessity of the second statement, i.e., $1 \Rightarrow 2$. If the allocation rule is implementable with outcome-contingent transfers, there exists T such that for all $\theta, \theta' \in \Theta$

$$\text{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T \geq 0.$$

For any $\lambda_{\theta\theta'} > 0$, we have

$$\lambda_{\theta\theta'}(\text{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T) \geq 0.$$

For any $\lambda \in \ker_+(D\pi)$, summing over all θ, θ', the second term is zero. What left is

$$\sum_{\theta,\theta'} \lambda_{\theta\theta'} \text{vl}(\theta, \theta') \geq 0.$$

Since this holds for all $\lambda \in \ker_+(D\pi)$, we have $\text{VL} \in [\ker_+(D\pi)]^*$.

The equivalence between statements 2 and 3 follows by a property of the convex envelope (see Boyd and Vandenberghe, 2004 p.119)

$$conv(D)(\mathbf{z}) = \inf\left\{\sum_{\theta,\theta'\in\Theta}\lambda_{\theta\theta'}\mathrm{vl}(\theta,\theta')\Big|\sum_{\theta\neq\theta'\in\Theta}\lambda_{\theta\theta'}=1,\lambda_{\theta\theta'}\geq 0,\sum_{\theta,\theta'\in\Theta}\lambda_{\theta\theta'}d\pi(\theta,\theta')=\mathbf{z}\right\},$$

$$conv(D)(\mathbf{0}) = \inf\left\{\lambda\cdot\mathrm{VL}\Big|\sum_{\theta\neq\theta'\in\Theta}\lambda_{\theta\theta'}=1,\lambda\in\ker_+(D\pi)\right\}.$$

Then $conv(D)(\mathbf{0}) \geq 0$ is equivalent to $\lambda\cdot\mathrm{VL}\geq 0$ for all $\lambda\in\ker_+(D\pi)$, which is $\mathrm{VL}\in[\ker_+(D\pi)]^*$

The sufficiency condition states that as long as the convex envelope intercept is weakly positive, the allocation rule is implementable with outcome-contingent transfers. Now we construct a transfer T such that for all $\theta,\theta'\in\Theta$, $\mathrm{vl}(\theta,\theta')+d\pi(\theta,\theta')\cdot T\geq 0$. Suppose $conv(D)(\mathbf{0})\geq 0$. Let T be the negative of any subgradient of $conv(D)(\cdot)$ at $d\pi=\mathbf{0}$, i.e.,

$$-T\in\partial conv(D)(\mathbf{0}),$$

where $\partial conv(D)(\mathbf{0})$ denotes the subdifferential of $conv(D)(\cdot)$ at $\mathbf{0}$. By the definition of the convex envelope and subgradient, for all $\theta,\theta'\in\Theta$,

$$\mathrm{vl}(\theta,\theta')\geq conv(D)(d\pi(\theta,\theta'))\geq conv(D)(\mathbf{0})-T\cdot d\pi(\theta,\theta').$$

$$\mathrm{vl}(\theta,\theta')+d\pi(\theta,\theta')\cdot T\geq conv(D)(\mathbf{0})\geq 0.$$

Geometrically, we rotate the difference set D around $(\mathbf{0},conv(D)(\mathbf{0}))$ such that the convex envelope is above the vl = 0 plane while preserving the intercept with the vl-axis. In the example in Figure 3.2, all deviation losses will be positive after the rotation, as shown in Figure 3.4.

We summarize the construction of the transfer.

Proposition 6. *If the allocation rule is implementable with outcome-contingent transfers, any $T\in-\partial conv(D)(\mathbf{0})$ can implement the allocation rule.*

The convex envelope intercept provides a measure of robustness of the implementation. Given the transfer identified above, the deviation loss is always above $conv(D)(\mathbf{0})$. (This is reminiscent of the definition of ϵ-Nash equilibrium.) Given this observation, we can characterize when an allocation rule is strictly implementable, i.e., the incentive to report truthfully is strict. We say that the allocation rule is *strictly implementable* if there exists a transfer T such that for all $\theta\neq\theta'\in\Theta$,

$$\pi(\theta)\cdot(v(\theta)+T)>\pi(\theta')\cdot(v(\theta)+T).$$

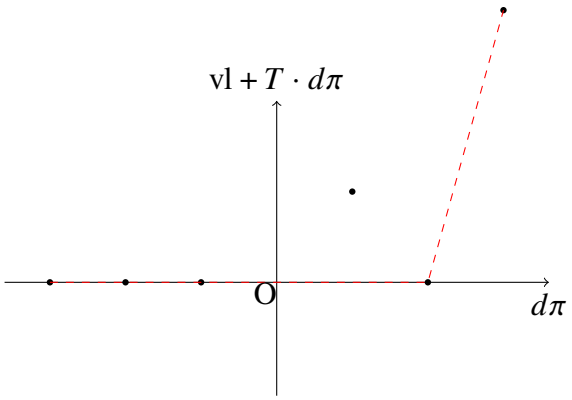


Figure 3.4: The Deviation Loss with Transfer

Proposition 7. *If the convex envelope intercept is strictly positive, the allocation rule is strictly implementable.*

Similar to the definition of ϵ-Nash equilibrium, we can also adopt a weaker condition on implementation. For any ϵ, we say that an allocation rule is *ϵ-implementable with outcome-contingent transfers* if there exists a transfer T such that the total gain from deviation is always less than ϵ, i.e., for all θ, $\theta' \in \Theta$,

$$\pi(\theta') \cdot (v(\theta) + T) - \pi(\theta) \cdot (v(\theta) + T) \leq \epsilon.$$

Corollary 5. *The allocation rule is $-conv(D)(\mathbf{0})$-implementable.*

3.4 Discussions

Rochet, 1987 studies the implementability condition with report-contingent transfer and shows that the cyclic monotonicity is sufficient and necessary. Formally, cyclic monotonicity is equivalent to the existence of a function $\tilde{t} : \Theta \to \mathbb{R}$ such that for all θ, θ',

$$\pi(\theta) \cdot v(\theta) + \tilde{t}(\theta) \geq \pi(\theta') \cdot v(\theta) + \tilde{t}(\theta').$$

Note the difference between report-contingent transfer and outcome-contingent transfer. Our implementation additionally requires that $\tilde{t}$ is linear in the allocation rule. If an allocation rule is implementable with outcome-contingent transfers, then the function $\tilde{t}$ must exist: $\tilde{t}(\theta) = \pi(\theta) \cdot T$. We thus obtain a necessary condition.

Observation 1. The allocation rule is implementable with outcome-contingent transfers only if $\mathrm{vl}(\cdot, \cdot)$ satisfies cyclic monotonicity, i.e., for all $\theta_1, \cdots, \theta_k \in \Theta$,

$$\mathrm{vl}(\theta_1, \theta_2) + \mathrm{vl}(\theta_2, \theta_3) + \cdots + \mathrm{vl}(\theta_k, \theta_1) \geq 0.$$

Yet the condition of cyclic monotonicity does not guarantee the existence of outcome-contingent transfers. In Example 2, a report-contingent transfer can implement the allocation rule. Thus, cyclic monotonicity holds. However, no outcome-contingent transfer can implement the allocation rule. Hence, our condition in Theorem 7 is stronger than cyclic monotonicity.

Special Cases

Our model imposes no assumptions on the allocation rule, the state space, or the valuation structure. Next, we shall investigate some special cases where we impose more structure on each of these model primitives. We first show that when $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly independent, cyclic monotonicity is also sufficient.

Proposition 8. *When $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly independent, the allocation rule is implementable with outcome-contingent transfers if and only if* $\mathrm{vl}(\cdot,\cdot)$ *satisfies cyclic monotonicity.*

Proof of Proposition 8. Observation 1 shows the necessity. We only need to show the sufficiency. By our previous discussion, cyclic monotonicity already ensures the existence of $\tilde{t}$. We only need to show that there exists a transfer $T \in \mathbb{R}^n$ such that for all $\theta \in \Theta$,

$$\pi(\theta) \cdot T = \tilde{t}(\theta).$$

We rewrite it in matrix form. We define $\mathbf{\Pi}$ be the $|\Theta| \times n$ matrix representing $\pi : \theta \to \Delta(X)$ and $\tilde{T}$ be the $|\Theta|$-dimension column vector representing $\tilde{t}(\cdot) : \Theta \to \mathbb{R}$. So the matrix form of the above linear system is

$$\mathbf{\Pi} T = \tilde{T}. \tag{3.1}$$

The vector $\tilde{T}$ lies in the span of column vectors of $\mathbf{\Pi}$. Thus, there exists a T satisfying the above matrix equation if and only if $rank(\mathbf{\Pi}) = rank(\mathbf{\Pi}, \tilde{T})$ where $\mathbf{\Pi}, \tilde{T}$ represents the augmented matrix. Since $\pi(\theta)$ is linearly independent, $rank(\mathbf{\Pi}) = |\Theta| = rank(\mathbf{\Pi}, \tilde{T})$. □

This proposition highlights the difference between allocation rules that are implementable with outcome-contingent transfers versus report-contingent transfers. If the cardinality of the outcomes is larger than the cardinality of states, then we have more flexibility in setting $t(x)$ to induce truthful reports. That is, generically, a matrix Π such that $|X| \geq |\Theta|$ guarantees the existence of a solution to Equation (3.1). Conversely, when $|X| < |\Theta|$, it is more likely that no solution exists. This

insight sheds light on the motivating examples. How far are the allocation rules that are implementable with outcome-contingent transfers compared to the ones with standard transfers? The answer largely lies in the granularity of outcomes versus states. In the political transfer example, as the state is high-dimensional, it is very hard to implement with outcome-contingent transfers if the evaluations are finite. However, the two implementation notions are closer given a larger evaluation set.

Next, we show that in some cases, it is easy to check whether an allocation rule is implementable. Consider the support problem in Figure 3.5. The outcome-contingent transfer rule is the common vector adjustment in the vector field that makes $v(\theta) + T$ (the outer normal of) a supporting hyperplane. Consider the point $\pi(\theta_4)$ in Figure 3.5. As $\pi(\theta_4)$ is in the convex hull of $\{\pi(\theta)|\theta \in \Theta\}$, no hyperplane can support $\{\pi(\theta)|\theta \in \Theta\}$. Therefore, the outer normal $v(\theta_4) + T$ must be zero and $T = -v(\theta_4)$.[7] Consequently, this is the only candidate transfer that we need to check. This insight carries over in general.

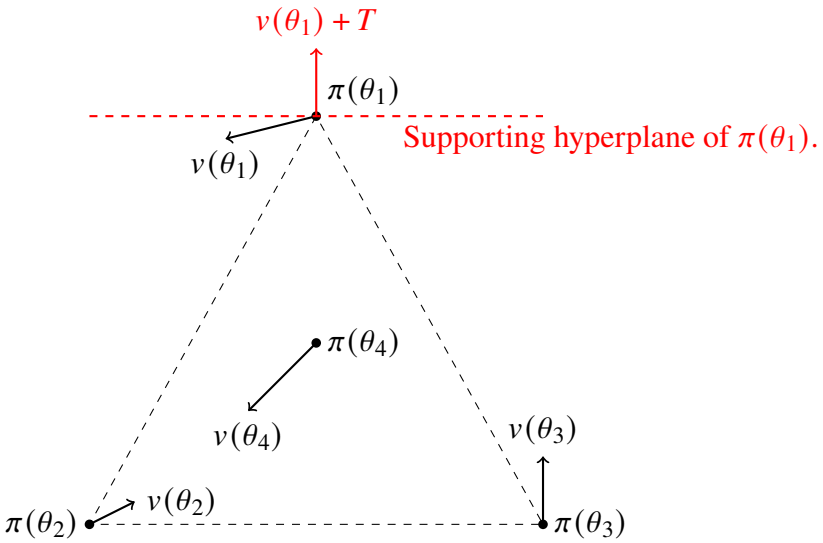


Figure 3.5: The Support Problem

Proposition 9. *Suppose that some $\pi(\theta_i)$ is in the interior of the convex hull of $\{\pi(\theta)|\theta \in \Theta\}$. The allocation rule π is implementable with outcome-contingent transfers if and only if $T = -v(\theta_i)$ implements π.*

We can apply this result to our Example 2. As $\pi(\theta_2)$ is in the interior of the convex hull of $\{\pi(\theta)|\theta \in \Theta\}$, it is without loss to consider only the transfer $T = -v(\theta_2)$.

[7] Formally, $v(\theta_4) + T$ can be non-zero but must be orthogonal to the affine hull of $\{\pi(\theta)|\theta \in \Theta\}$. But then it is without loss to consider only $T = -v(\theta_4)$.

But this transfer cannot elicit θ_3 to report truthfully. Thus, the allocation rule is not implementable.

Note that this proposition has no bite if the points $\{\pi(\theta)|\theta \in \Theta\}$ are in a convex position (also known as convex independent). On the other hand, when $\{\pi(\theta)|\theta \in \Theta\}$ are convex dependent, implementation is generally difficult. Even when such an allocation can be implemented, the transfer rule is very restricted. In Appendix 3.7, we consider the planner's design problem where he optimizes over allocation-transfer rules. We show that for a general objective function, it is without loss to restrict attention to convex independent allocation rules.

In addition, we show that when the state space is small, cyclic monotonicity is also sufficient.

Proposition 10. *When* $|\Theta| \leq 3$*, the allocation rule is implementable with outcome-contingent transfers if and only if* $\mathrm{vl}(\cdot,\cdot)$ *satisfies cyclic monotonicity.*

By Proposition 8, the conclusion follows when $\pi(\theta)$ are linearly independent. Now suppose $\pi(\theta)$ are linearly dependent. When $|\Theta| = 2$, linear dependence of $\pi(\theta)$ implies $\pi(\theta_1) = \pi(\theta_2)$ and the conclusion holds trivially. When $|\Theta| = 3$, the linear dependence of $\pi(\theta)$ is equivalent to convex dependence. Suppose that $\pi(\theta_1)$ is a convex combination of $\pi(\theta_2)$ and $\pi(\theta_3)$. By Proposition 9, it is without loss to take $T = -v(\theta_1)$. Given this transfer, all the incentive compatibility constraints reduce to weak monotonicity.[8] Thus, cyclic monotonicity is also sufficient.

The argument above fails catastrophically for $|\Theta| \geq 4$. First, when there are more than three states, linear dependence does not imply convex dependence. Second, even if convex dependence of $\{\pi(\theta)|\theta \in \Theta\}$ holds, we can no longer reduce all IC constraints to weak monotonicity. This occurs in Example 2, where $|\Theta| = 4$ and

[8]To see this, let $\pi(\theta_1) = \lambda\pi(\theta_2) + (1-\lambda)\pi(\theta_3)$. Given $T = -v(\theta_1)$, the incentive compatibility requires that for all $i \neq j \in \{1,2,3\}$,

$$\pi(\theta_i) \cdot (v(\theta_i) - v(\theta_1)) \geq \pi(\theta_j) \cdot (v(\theta_i) - v(\theta_1)).$$

When $i = 1$, the inequality holds trivially. When $j = 1$, the inequality reduces to weak monotonicity between θ_i and θ_1. When $i, j \in \{2,3\}$, replacing $\pi(\theta_j)$ with

$$\frac{\pi(\theta_1) - \lambda_i\pi(\theta_i)}{1-\lambda_i},$$

where $\lambda_i = \lambda$ if $i = 2$ and $\lambda_i = 1-\lambda$ otherwise, the inequality also reduces to weak monotonicity between θ_i and θ_1.

cyclic monotonicity holds. Thus, when $|\Theta| \geq 4$, cyclic monotonicity may no longer be sufficient.

In the incentive compatibility constraint, we take expectation of the random outcome. Thus, we can view the implementation with outcome-contingent transfers as an interim condition. A more demanding notion can require the allocation rule to be ex-post implementable with outcome-contingent transfers, i.e., if there exists a transfer t such that for all θ and $x \in \mathbf{supp}\{\pi(\cdot|\theta)\}$,

$$v(\theta, x) + t(x) \geq v(\theta, x') + t(x'), \forall x' \in X.$$

It turns out that this ex-post implementability is equivalent to the following sufficient condition on the valuation structure.[9]

Proposition 11. *The allocation rule is implementable with outcome-contingent transfers if for any sequence of* $(\theta_1, x_1), \cdots, (\theta_m, x_m), (\theta_{m+1}, x_{m+1}) = (\theta_1, x_1)$ *where* $x_i \in \boldsymbol{supp}\{\pi(\cdot|\theta_i)\}$,

$$\sum_{i=1}^{m} v(\theta_i, x_i) \geq \sum_{i=1}^{m} v(\theta_i, x_{i+1}).$$

Another important case is when the agent's preference is separable in the state and outcome. We say that the agent's preference is *additively separable* if there exists v_1 and v_2 such that

$$v(\theta, x) = v_1(\theta) + v_2(x).$$

It includes the case where the agent's preference is state-independent. State-independent preference, or transparent motive, has been widely studied in the communication and persuasion literature (see, for example, Chakraborty and Harbaugh, 2010; Lipnowski and Ravid, 2020; Lipnowski, Ravid, and Shishkin, 2022). When the agent's preference is additively separable, the transfer $t(x) = -v_2(x)$ can implement all allocation rules. Moreover, the converse is true as well, i.e., this is the only preference where all allocation rules are implementable with outcome-contingent transfers.

3.5 Conclusion

We study whether we can implement a randomized allocation rule with outcome-contingent transfers. For this implementation, we characterize sufficient and necessary conditions. One natural extension is to study a principal's revenue maximization

[9] We thank one anonymous referee for providing this result.

problem when the agent reports through a noisy signal, which can be viewed as our allocation rule. The principal allocates one indivisible item conditional on the message generated by the signal. Given the agent's IR constraint, the principal designs transfers to maximize revenue. We leave this question to future research.

3.6 Omitted Proofs

Proof of Theorem 7. $1 \Rightarrow 2$. If the allocation rule is implementable with outcome-contingent transfers, there exists T such that for all $\theta, \theta' \in \Theta$

$$\mathrm{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T \geq 0.$$

For any $\lambda_{\theta\theta'} > 0$, we have

$$\lambda_{\theta\theta'}(\mathrm{vl}(\theta, \theta') + d\pi(\theta, \theta') \cdot T) \geq 0.$$

For any $\lambda \in \ker_+(D\pi)$, summing over all θ, θ',

$$\sum_{\theta,\theta'} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') \geq 0$$

we have $\mathrm{VL} \in [\ker_+(D\pi)]^*$.

$2 \Rightarrow 3$. If $\mathrm{VL} \in [\ker_+(D\pi)]^*$, the optimal value of following linear programming problem is zero.

$$\begin{aligned} \min & \sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') \\ s.t. & \sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} d\pi(\theta, \theta') = 0, \\ & \lambda_{\theta\theta'} \geq 0. \end{aligned} \tag{3.2}$$

Then, for any $\lambda_{\theta\theta'} \geq 0$ satisfies $\sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} d\pi(\theta, \theta') = 0$ and $\sum_{\theta \neq \theta' \in \Theta} \lambda_{\theta\theta'} = 1$, we have $\sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') \geq 0$.

By the definition of convex envelope,[10]

$$conv(D)(\mathbf{z}) = \inf\{ \sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') \Big| \sum_{\theta \neq \theta' \in \Theta} \lambda_{\theta\theta'} = 1, \lambda_{\theta\theta'} \geq 0, \sum_{\theta,\theta' \in \Theta} \lambda_{\theta\theta'} d\pi(\theta, \theta') = \mathbf{z} \}.$$

Thus we get $conv(D)(\mathbf{0}) \geq 0$.

$3 \Rightarrow 1$. Since $conv(D)(0) \geq 0$, the convex hull of set D, $conhull(D)$, and the convex set $\{(\mathbf{0}, l) \big| l < 0, \mathbf{0} \in \mathbb{R}^n\}$ have no intersection. By Separating Hyperplane Theorem, there exists a non-zero vector $(\bar{T}, \alpha)$ where $\bar{T} \in \mathbb{R}^n, \alpha \geq 0$ such that for any $(d\pi(\theta, \theta'), \mathrm{vl}(\theta, \theta')) \in D$ and $l < 0$ we have

$$d\pi(\theta, \theta') \cdot \bar{T} + \alpha \mathrm{vl}(\theta, \theta') > \alpha l. \tag{3.3}$$

[10] Here we adopt the convention that $\inf \emptyset = +\infty$.

If $\alpha = 0$, we get $d\pi(\theta, \theta') \cdot \bar{T} > 0$ and $d\pi(\theta', \theta) \cdot \bar{T} > 0$. But $d\pi(\theta, \theta') + d\pi(\theta', \theta) = 0$, a contradiction. Then it must be that $\alpha > 0$. Set $T = \frac{\bar{T}}{\alpha}$, then by (3.3)

$$d\pi(\theta, \theta') \cdot T + \text{vl}(\theta, \theta') > l.$$

As $l < 0$, take the supremum of l,

$$d\pi(\theta, \theta') \cdot T + \text{vl}(\theta, \theta') \geq 0.$$

This implies that T is the transfer that implements the allocation rule. □

Proof of Proposition 9. The "if" part is obvious and we prove the "only if" part. Suppose that the allocation rule $\{\pi(\theta)\}$ is implementable. We assume that the transfer T' implements this allocation rule. Since $\pi(\theta_i)$ is in the interior of the convex hull of $\{\pi(\theta) | \theta \in \Theta\}$, there exists $\lambda(\theta) > 0$ such that $\sum_{\theta \neq \theta_i} \lambda(\theta) = 1$ and $\pi(\theta_i) = \sum_{\theta \neq \theta_i} \lambda(\theta)\pi(\theta)$.

By the incentive-compatible constraint, we have that for any $\theta \neq \theta_i$,

$$\lambda(\theta)\pi(\theta_i) \cdot (v(\theta_i) + T') \geq \lambda(\theta)\pi(\theta) \cdot (v(\theta_i) + T').$$

Sum them up, we get

$$\pi(\theta_i) \cdot (v(\theta_i) + T') \geq \pi(\theta_i) \cdot (v(\theta_i) + T').$$

Consequently, all above inequalities must be equalities

$$\pi(\theta_i) \cdot (v(\theta_i) + T') = \pi(\theta) \cdot (v(\theta_i) + T')$$

for all $\theta \neq \theta_i$.

Next, we verify that $T = -v(\theta_i)$ also implements the allocation rule. For any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
\pi(\theta) \cdot (v(\theta) + T) &= \pi(\theta) \cdot (v(\theta) + T') - \pi(\theta) \cdot (v(\theta_i) + T') \\
&\geq \pi(\theta') \cdot (v(\theta) + T') - \pi(\theta_i) \cdot (v(\theta_i) + T') \\
&= \pi(\theta') \cdot (v(\theta) + T') - \pi(\theta') \cdot (v(\theta_i) + T') \\
&= \pi(\theta') \cdot (v(\theta) + T).
\end{aligned}$$

Thus $T = -v(\theta_i)$ implements the allocation rule $\{\pi(\theta)\}_{\theta \in \Theta}$. □

Proof of Proposition 10. Observation 1 shows the necessity. We only need to show sufficiency. When $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly independent, the conclusion holds by Proposition 8. Now suppose $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly dependent.

When $|\Theta| = 1$, the problem is trivial. When $\Theta = \{\theta_1, \theta_2\}$, the only linearly dependent case is $\pi(\theta_1) = \pi(\theta_2)$. It trivially satisfies the cyclic monotonicity condition and the allocation rule is implementable.

Now suppose $\Theta = \{\theta_1, \theta_2, \theta_3\}$. The result trivially holds when $\pi(\theta_1) = \pi(\theta_2) = \pi(\theta_3)$. For the other cases, there is a unique $t \in [0, 1]$ such that

$$\pi(\theta_3) = t\pi(\theta_1) + (1-t)\pi(\theta_2) \tag{3.4}$$

and $\pi(\theta_1) \neq \pi(\theta_2)$. This holds without loss of generality, since $\{\pi(\theta)\}_{\theta\in\Theta}$ are linearly dependent and $\pi(\theta) \geq \mathbf{0}$. Consequently, the dimension of $\ker_+(D\pi)$ is 1. For any $\lambda \in \ker_+(D\pi)$, the coefficient of $\pi(\theta_i)$ in $\sum_{\theta,\theta'\in\Theta} \lambda_{\theta\theta'} d\pi(\theta, \theta')$ is

$$\sum_{j\neq i} (\lambda_{\theta_i\theta_j} - \lambda_{\theta_j\theta_i}).$$

Since the dimension of the kernel space is 1, by (3.4), there exists a real number K such that

$$\begin{aligned} \sum_{j\neq 1} (\lambda_{\theta_1\theta_j} - \lambda_{\theta_j\theta_1}) &= Kt \\ \sum_{j\neq 2} (\lambda_{\theta_2\theta_j} - \lambda_{\theta_j\theta_2}) &= K(1-t) \\ \sum_{j\neq 3} (\lambda_{\theta_3\theta_j} - \lambda_{\theta_j\theta_3}) &= -K. \end{aligned} \tag{3.5}$$

Take any $\lambda \in \ker_+(D\pi)$. If there exists a cycle $(s_1, s_2, \cdots, s_k) \subseteq \Theta$ such that $\lambda_{s_1s_2} \times \cdots \times \lambda_{s_\kappa s_1} \neq 0$. Then let

$$y = \min\{\lambda_{s_1s_2}, \cdots, \lambda_{s_\kappa s_1}\} > 0,$$

and update the values of $\lambda_{s_1s_2}, \lambda_{s_2s_3}, \cdots, \lambda_{s_\kappa s_1}$ as the following:

$$\begin{aligned} \lambda_{s_1s_2} &\leftarrow \lambda_{s_1s_2} - y \\ &\cdots \\ \lambda_{s_\kappa s_1} &\leftarrow \lambda_{s_\kappa s_1} - y. \end{aligned}$$

Let λ^* denote the updated value. λ^* still satisfies equation (3.5). This implies $\lambda^* \in \ker_+(D\pi)$.

$$\sum_{\theta\theta'} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') - \sum_{\theta\theta'} \lambda^*_{\theta\theta'} \mathrm{vl}(\theta, \theta') = y \sum_{i=1}^{\kappa} \mathrm{vl}(s_i, s_{i+1}) \geq 0$$

by cyclic monotonicity. So it suffices to show that for all $\lambda^* \in \ker_+(D\pi)$, we have $\lambda^* \cdot \mathrm{VL} \geq 0$.

Thus, we can assume that there is no cycle that $(s_1, s_2, \cdots, s_k)$ such that $\lambda_{s_1 s_2} \times \cdots \times \lambda_{s_\kappa s_1} \neq 0$. As $|\Theta| = 3$, there must be a θ_i such that

$$\lambda_{\theta_j \theta_i} = 0, \ \forall j \neq i.$$

We say that such i has the *lowest topological order*. And there must be a θ_i such that

$$\lambda_{\theta_i \theta_j} = 0, \ \forall j \neq i.$$

We say that such i has the *highest topological order*. We consider two cases.

Case 1: $K \geq 0$. The lowest topological order index i must be 1 or 2. By symmetry, we assume that it is 1. And the highest topological order index must be 3. Then by (3.5),

$$\begin{aligned} \lambda_{\theta_1\theta_2} + \lambda_{\theta_1\theta_3} &= Kt, \\ \lambda_{\theta_2\theta_3} - \lambda_{\theta_1\theta_2} &= K(1-t). \end{aligned} \tag{3.6}$$

We calculate $\mathrm{VL} \cdot \lambda$,

$$\begin{aligned} \sum_{\theta\theta'} \lambda_{\theta\theta'} \mathrm{vl}(\theta, \theta') &= \lambda_{\theta_1\theta_2} \mathrm{vl}(\theta_1, \theta_2) + \lambda_{\theta_1\theta_3} \mathrm{vl}(\theta_1, \theta_3) + \lambda_{\theta_2\theta_3} \mathrm{vl}(\theta_2, \theta_3) \\ &= \lambda_{\theta_1\theta_2} \mathrm{vl}(\theta_1, \theta_2) + \lambda_{\theta_1\theta_3} [\pi(\theta_1) - \pi(\theta_3)] \cdot v(\theta_1) + \lambda_{\theta_2\theta_3} [\pi(\theta_2) - \pi(\theta_3)] \cdot v(\theta_2) \\ &= \lambda_{\theta_1\theta_2} \mathrm{vl}(\theta_1, \theta_2) + (1-t)\lambda_{\theta_1\theta_3} \mathrm{vl}(\theta_1, \theta_2) + t\lambda_{\theta_2\theta_3} \mathrm{vl}(\theta_2, \theta_1) \\ &= Kt(1-t)(\mathrm{vl}(\theta_1, \theta_2) + \mathrm{vl}(\theta_2, \theta_1)) + t\lambda_{\theta_1\theta_2}(\mathrm{vl}(\theta_1, \theta_2) + \mathrm{vl}(\theta_2, \theta_1)) \\ &\geq 0, \end{aligned}$$

where the third equality follows by replacing $\pi(\theta_3)$ with $t\pi(\theta_1) + (1-t)\pi(\theta_2)$, and the last equality follows by (3.6).

Case 2: $K < 0$. The highest topological order index i must be 1 or 2. By symmetry, we assume that it is 1. And the lowest topological order index must be 3. Then by (3.5),

$$\lambda_{\theta_2\theta_1} + \lambda_{\theta_3\theta_1} = -Kt,$$

$$\lambda_{\theta_3\theta_2} - \lambda_{\theta_2\theta_1} = -K(1-t).$$

If $t = 0$, then $\lambda_{\theta_2\theta_1} = \lambda_{\theta_3\theta_1} = 0$ and $\text{vl}(\theta_3, \theta_2) = 0$, the value $\sum_{\theta\theta'} \lambda_{\theta\theta'}\text{vl}(\theta, \theta') = 0$. If $t > 0$, then

$$\begin{aligned}\sum_{\theta\theta'} \lambda_{\theta\theta'}\text{vl}(\theta, \theta') &= \lambda_{\theta_3\theta_1}\text{vl}(\theta_3, \theta_1) + \lambda_{\theta_3\theta_2}\text{vl}(\theta_3, \theta_2) + \lambda_{\theta_2\theta_1}\text{vl}(\theta_2, \theta_1) \\ &= \frac{t-1}{t}\lambda_{\theta_3\theta_1}\text{vl}(\theta_3, \theta_2) + \lambda_{\theta_3\theta_2}\text{vl}(\theta_3, \theta_2) + \frac{1}{t}\lambda_{\theta_2\theta_1}\text{vl}(\theta_2, \theta_3) \\ &= \frac{\lambda_{\theta_2\theta_1}}{t}(\text{vl}(\theta_2, \theta_3) + \text{vl}(\theta_3, \theta_2)) \\ &\geq 0.\end{aligned}$$

Hence, we have $\text{VL} \in [\ker_+(D\pi)]^*$. By Theorem 7, the allocation rule is implementable with outcome-contingent transfers. □

Proof of Proposition 11. By Kantorovich Duality (Theorem 5.10 in Villani et al., 2009), there exists $t : X \to \mathbb{R}$ such that for all for all θ and $x \in \textbf{supp}\{\pi(\cdot|\theta)\}$,

$$v(\theta, x) + t(x) \geq v(\theta, x') + t(x'), \forall x' \in X$$

if and only if $\lambda^* = \mu_0(\theta)\pi(x|\theta)$ is optimal solution for the following optimal transport problem,

$$\max_{\lambda \in \Delta(\Theta \times X)} \sum_{\theta, x} \lambda(\theta, x) v(\theta, x)$$
$$s.t. \lambda_\theta = \mu_0, \lambda_X = \nu,$$

where μ_0 is a full-support distribution on Θ and $\nu(x) = \sum_{\theta \in \Theta} \mu(\theta)\pi(x|\theta)$. Again by Theorem 5.10 in Villani et al., 2009, λ^* is the solution of above optimal transport problem if and only if for any sequence $(\theta_1, x_1), \cdots, (\theta_m, x_m), (\theta_{m+1}, x_{m+1}) = (\theta_1, x_1)$ where $(\theta_i, x_i) \in \textbf{supp}\{\lambda^*\}$,

$$\sum_{i=1}^{m} v(\theta_i, x_i) \geq \sum_{i=1}^{m} v(\theta_i, x_{i+1}).$$

Note that $(\theta_i, x_i) \in \textbf{supp}\{\lambda^*\}$ if and only if $x_i \in \textbf{supp}\{\pi(\cdot|\theta_i)\}$. □

Proof of Claim: All allocation rules are implementable with outcome-contingent transfers if and only if the agent's preference is additively separable.

The "if" part is taken care of by transfer $t(x) = -v_2(x)$. The "only if" part: Suppose all allocation rules are implementable with outcome-contingent transfers. Then we know that for any $\{\pi(\theta)\}_{\theta \in \Theta}$, by Observation 1, $\text{vl}(\cdot, \cdot)$ satisfies the cyclic

monotonicity condition. Then for any $\theta \neq \theta' \in \Theta, x \neq x' \in X$. If we consider $\pi(x|\theta) = 1, \pi(x'|\theta') = 1$, the cyclic monotonicity condition requires that

$$v(\theta, x) + v(\theta', x') \geq v(\theta', x) + v(\theta, x').$$

If we consider $\pi(x'|\theta) = 1, \pi(x|\theta') = 1$, the cyclic monotonicity condition requires that

$$v(\theta, x) + v(\theta', x') \leq v(\theta', x) + v(\theta, x').$$

Then we know that for any $\theta \neq \theta' \in \Theta, x \neq x' \in X$, we must have $v(\theta, x) - v(\theta, x') = v(\theta', x) - v(\theta', x')$. Then fix $x_0 \in X$, then there is $v_2 \colon X \to \mathbb{R}$ such that $v(\theta, x) - v(\theta, x_0) = v_2(x)$ for all $\theta \in \Theta$. Thus we let $v_1(\theta) = v(\theta, x_0)$, then $v(\theta, x) = v_1(\theta) + v_2(x)$ which implies the agent's preference is additive separable. □

3.7 Optimization over Allocation Rules

In this section, we take allocation rules as endogenous and consider a design problem. The planner's ex-post payoff function is $f(\theta, x, t)$. The planner sets up an outcome-contingent allocation and transfer rule $(\pi(\theta), T)$ to maximize expected payoff

$$E_\theta\{\sum_x \pi(x|\theta) f(x, \theta, t(x))\}$$

subject to the IC constraint

$$\forall \theta, \theta' \in \Theta, \quad \pi(\theta) \cdot (v(\theta) + T) \geq \pi(\theta') \cdot (v(\theta) + T)$$

and IR (participation) constraint

$$\forall \theta \in \Theta, \quad \pi(\theta) \cdot (v(\theta) + T) \geq 0.$$

We show that it is without loss to restrict attention to convex independent allocation rules.

Proposition 12. *It is without loss for the planner to focus on convex independent allocation rules.*

Proof. Suppose that (π, T) satisfies the IC and IR constraints. We show that there is a convex independent allocation rule π' such that (π', T) yields a weakly larger payoff for the planner.

Let $\Theta' \subset \Theta$ collect all θ such that $\pi(\theta)$ is the extreme point of the convex hull of $\{\pi(\theta)|\theta \in \Theta\}$. Fix some $\theta_i \in \Theta$. There exists $\lambda(\cdot) : \Theta' \to \mathbb{R}_{\geq 0}$ such that

$$\pi(\theta_i) = \sum_{\theta \in \Theta'} \lambda(\theta)\pi(\theta) \quad \text{and} \quad \sum_{\theta \in \Theta'} \lambda(\theta) = 1.$$

Note that the planner's expected payoff conditional on θ_i is

$$\sum_x \pi(x|\theta_i) f(x, \theta_i, t(x)) = \sum_{\theta \in \Theta'} \lambda(\theta) \sum_x \pi(x|\theta) f(x, \theta_i, t(x)).$$

There must exist some $\theta_i' \in \Theta'$ such that

$$\sum_x \pi(x|\theta_i') f(x, \theta_i, t(x)) \geq \sum_x \pi(x|\theta_i) f(x, \theta_i, t(x)).$$

We define a new allocation rule π' by $\pi'(\theta_i) = \pi(\theta_i')$. Note that (π', T) generates a weakly higher payoff for the planner. Lastly, by Proposition 9, agent θ_i's payoff does not change,

$$(v(\theta_i) + T) \cdot \pi(\theta_i) = (v(\theta_i) + T) \cdot \pi'(\theta_i).$$

Thus, IR still holds. The set of IC constraints is smaller due to

$$\{\pi'(\theta) | \theta \in \Theta\} = \{\pi(\theta) | \theta \in \Theta'\} \subset \{\pi(\theta) | \theta \in \Theta\}.$$

Thus, IC still holds.

□

BIBLIOGRAPHY

Archer, Aaron and Robert Kleinberg (2008). "Truthful germs are contagious: a local to global characterization of truthfulness". In: *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 21–30.

Ashlagi, Itai et al. (2010). "Monotonicity and implementability". In: *Econometrica* 78.5, pp. 1749–1772.

Bergemann, Dirk, Stephen Morris, and Satoru Takahashi (2012). "Efficient auctions and interdependent types". In: *American Economic Review Papers & Proceedings* 102.3, pp. 319–324.

Bikhchandani, Sushil et al. (2006). "Weak monotonicity characterizes deterministic dominant-strategy implementation". In: *Econometrica* 74.4, pp. 1109–1132.

Boyd, Stephen P and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Carbajal, Juan Carlos and Rudolf Müller (2015). "Implementability under monotonic transformations in differences". In: *Journal of Economic Theory* 160, pp. 114–131.

– (2017). "Monotonicity and revenue equivalence domains by monotonic transformations in differences". In: *Journal of Mathematical Economics* 70, pp. 29–35.

Carroll, Gabriel (2012). "When are local incentive constraints sufficient?" In: *Econometrica* 80.2, pp. 661–686.

Chakraborty, Archishman and Rick Harbaugh (2010). "Persuasion by cheap talk". In: *American Economic Review* 100.5, pp. 2361–2382.

Doval, Laura and Vasiliki Skreta (2022). "Mechanism design with limited commitment". In: *Econometrica* 90.4, pp. 1463–1500.

Frongillo, Rafael M and Ian A Kash (2021). "General truthfulness characterizations via convex analysis". In: *Games and Economic Behavior* 130, pp. 636–662.

Gui, Hongwei, Rudolf Müller, and Rakesh V Vohra (2004). *Dominant strategy mechanisms with multidimensional types*. Tech. rep. Discussion Paper.

Jehiel, Philippe, Benny Moldovanu, and Ennio Stacchetti (1999). "Multidimensional mechanism design for auctions with externalities". In: *Journal of economic theory* 85.2, pp. 258–293.

Kushnir, Alexey I and Lev V Lokutsievskiy (2021). "When is a monotone function cyclically monotone?" In: *Theoretical Economics* 16.3, pp. 853–879.

Lin, Xiao and Ce Liu (2024). "Credible Persuasion". In: *Journal of Political Economy* 132.7, pp. 2228–2273.

Lipnowski, Elliot and Doron Ravid (2020). "Cheap talk with transparent motives". In: *Econometrica* 88.4, pp. 1631–1660.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin (2022). "Persuasion via weak institutions". In: *Journal of Political Economy* 130.10, pp. 2705–2730.

McAfee, R Preston and John McMillan (1988). "Multidimensional incentive compatibility and mechanism design". In: *Journal of Economic theory* 46.2, pp. 335–354.

Müller, Rudolf, Andrés Perea, and Sascha Wolf (2007). "Weak monotonicity and Bayes–Nash incentive compatibility". In: *Games and Economic Behavior* 61.2, pp. 344–358.

Myerson, Roger B (1981). "Optimal auction design". In: *Mathematics of Operations Research* 6.1, pp. 58–73.

Nguyen, Anh and Teck Yong Tan (2021). "Bayesian persuasion with costly messages". In: *Journal of Economic Theory* 193, p. 105212. ISSN: 0022-0531.

Perez-Richet, Eduardo and Vasiliki Skreta (2022). "Test design under falsification". In: *Econometrica* 90.3, pp. 1109–1142.

Roberts, Kevin (1979). "The characterization of implementable choice rules". In: *Aggregation and revelation of preferences* 12.2, pp. 321–348.

Rochet, Jean-Charles (1987). "A necessary and sufficient condition for rationalizability in a quasi-linear context". In: *Journal of Mathematical Economics* 16.2, pp. 191–200.

Saks, Michael and Lan Yu (2005). "Weak monotonicity suffices for truthfulness on convex domains". In: *Proceedings of the 6th ACM conference on Electronic commerce*, pp. 286–293.

Tadelis, Steve and Ilya Segal (2005). *Lectures in Contract Theory*.

Villani, Cédric et al. (2009). *Optimal transport: old and new*. Vol. 338. Springer.