EC 2010B: Microeconomic Theory II

Michael Powell

March, 2018

Contents

Disclaimer

These lecture notes were written for a first-year Ph.D. course on General Equilibrium Theory and Contract Theory. They are a work in progress and may therefore contain errors. The GE section borrows liberally from Levin's (2006) notes and from Wolitzky's (2016) notes. I am grateful to Angie Acquatella for many helpful comments. Any comments or suggestions would be greatly appreciated.

Introduction

In the first three weeks of this course, our goal is to develop a parsimonious model of the overall economy to study the interaction of individual consumers and firms in perfectly competitive *decentralized* markets. The resulting framework has provided the workhorse micro-foundations for much of modern macroeconomics, international trade, and financial economics. In the last three and a half weeks, we will begin to study *managed* transactions and in particular how individuals should design institutions such as contracts and property rights allocations to governance structures, to achieve desirable outcomes.

Part I

The Invisible Hand

The main ideas of general equilibrium theory have a long history, going back to Adam Smith's evocative descriptions of how competition channels individual self-interest in the social interest and how a sense of "coherence among the vast numbers of individuals and seemingly separate decisions" (Arrow, 1972) can arise in the economy without explicit design. General equilibrium theory addresses how this aggregate "coherence" emerges from individual interactions and can potentially lead to socially desirable allocations of goods and services in the economy. The mechanism through which this coherence emerges is, of course, the price mechanism. Individuals facing the same, suitably determined, prices will end up making decisions that are well-coordinated at the economy-wide level.

What distinguishes general equilibrium theory from partial equilibrium theory, which you have studied in Economics 2010a, is the idea that if we want to develop a theory of the price system for the economy as a whole, we have to consider the equilibrium in all markets in the economy simultaneously. As you can imagine, thinking about all markets simultaneously can be a complicated endeavor, since markets are interdependent: the price of computer chips will affect the price of software, cars, appliances, and so on. General equilibrium theory was the most active research area in economic theory for a good part of the 20th century and is therefore a very rich topic. Our goal over the first three weeks of the course is to cover only the basics.

In many ways, the first part of this class will be structured the way an applied theory paper is structured. We will start by talking about the setup of a model: who are the players, what do they do, what do they know, what are their preferences, what is the solution concept we will be using? Then we will partially characterize its solution, focusing on its efficiency properties in particular. What do I mean by partially characterize? I mean that we will be describing some properties of equilibria that we can talk about without actually solving the full model. That should take us through the first week.

In the second week, we will do a little bit of heavier lifting and begin with a question that is often easy to overlook but is very important to answer: does an equilibrium exist? Existence proofs can sometimes seem like a bit of an esoteric detour, but a good existence proof is especially useful if it helps build tools to answer other important questions about the model. After we establish that an equilibrium exists, we will ask questions like: when is there a unique equilibrium? Are equilibria stable? What are the testable implications of equilibrium behavior?

In the third week, we will put the model's solution concept, competitive equilibrium, on firmer microfoundations. A lot of economics is buried in the solution concept, and developing microfoundations for the solution concept is a useful way to flesh out some of the key insights. We will conclude with what I will blandly call "extensions," but are really the meat and potatoes of how general equilibrium theory gets used in practice. We will show how we can introduce firms, time, and uncertainty into the framework, and we will talk about how and when the main results we identified above apply in these settings as well.

Chapter 1

Pure Exchange Economies

A general equilibrium model describes three basic activities that take place in the economy: production, exchange, and consumption. For the first two weeks, we will set production aside and focus on the minimal number of modeling ingredients necessary to give you a flavor of the powerful results of general equilibrium theory.

We begin with a description of the model we will be using. As always, the exposition of an economic model specifies a complete description of the economic environment (players, actions, and preferences) and the solution concept that will be used to derive prescriptions and predictions.

The Model Formally, a pure exchange economy is an economy in which there are no production opportunities. There are I **consumers** $i \in \mathcal{I} = \{1, \ldots, I\}$ who buy, sell, and consume L **commodities**, $l \in \mathcal{L} = \{1, \ldots, L\}$.

A **consumption bundle for consumer** i is a vector $x_i = (x_{1,i}, \ldots, x_{L,i})$, where $x_i \in \mathcal{X}_i$, which is consumer i's **consumption set** and just describes her feasible consumption bundles. We will assume throughout that $\mathcal{X}_i$ contains the 0 vector and is a convex set. Consumer i has an **endowment** of the L commodities, which is described by a vector $\omega_i = (\omega_{1,i}, \ldots, \omega_{L,i})$ and has preferences over consumption bundles, which we assume can be represented by a utility function $u_i : \mathcal{X}_i \to \mathbb{R}$. A **pure exchange economy** is therefore a set $\mathcal{E} = \left((u_i, \omega_i)_{i \in \mathcal{I}}\right)$, which fully describes the model's primitives: the set of players, their preferences, and their endowments.

Each consumer takes **prices** $p = (p_1, \ldots, p_L)$ as given and solve her **consumer maximization problem**:

$$\max_{x_i \in \mathcal{X}_i} u_i(x_i) \text{ s.t. } p \cdot x_i \leq p \cdot \omega_i,$$

where the right-hand side of the consumer's budget constraint is her wealth, as measured by the market value of her endowment at prices p. Consumer i's feasible actions are therefore $x_i \in \mathcal{B}_i(p) \equiv \{x_i \in \mathcal{X}_i : p \cdot x_i \leq p \cdot \omega_i\}$, where we refer to $\mathcal{B}_i(p)$ as her **budget set at prices** p. Given prices p and endowment ω_i, we will refer to consumer i's optimal choices as her **Marshallian demand correspondence** and denote it by $x_i(p, p \cdot \omega_i)$. Simply put, all consumers do in this model is to choose their favorite consumption bundles in their budget sets.

We have now described the players and their actions, but no model description is complete without a solution concept. Here, the solution concept will be a Walrasian equilibrium, which will specify a set of prices and a consumption bundle for each consumer (we will refer to a collection of consumption bundles for each consumer as an **allocation**) that satisfy two properties: consumer optimization and market-clearing. Given prices p, each consumer optimally chooses her consumption bundle, and total demand for each commodity equals total supply.

Definition 1. A **Walrasian equilibrium** for the pure exchange economy $\mathcal{E}$ is a vector $\left(p^*, (x_i^*)_{i\in\mathcal{I}}\right)$ that satisfies:

1. Consumer optimization: for all consumers $i \in \mathcal{I}$,

$$x_i^* \in \operatorname{argmax}_{x_i \in \mathcal{B}_i(p^*)} u_i\left(x_i\right),$$

2. Market-clearing: for all commodities $l \in \mathcal{L}$,

$$\sum_{i\in\mathcal{I}} x_{l,i}^* = \sum_{i\in\mathcal{I}} \omega_{l,i}.$$

We have now fully specified the model, but before we start to go into more detail discussing the properties of Walrasian equilibria, there are a couple more important definitions to introduce. The first is the notion of a feasible allocation, which is just a collection of consumption bundles for each consumer for which the total amount consumed for each commodity does not exceed the total endowment of that commodity.

Definition 2. An allocation $(x_i)_{i \in \mathcal{I}} \in \mathbb{R}_+^{L \cdot I}$ is **feasible** if for all $l \in \mathcal{L}$, $\sum_{i \in \mathcal{I}} x_{l,i} \leq \sum_{i \in \mathcal{I}} \omega_{l,i}$.

The next definition is going to describe how we will be thinking about optimality in the economy. For a lot of optimization problems you have seen in your other courses, the appropriate notion of optimality is straightforward. For example, if a consumer has a well-defined utility function, it is straightforward to think about what is optimal for her given her budget set. Once we start thinking about environments with more than one consumer, we would, in some sense, like to maximize multiple objective functions (i.e., each consumer's utility) simultaneously. In general, there are no allocations that simultaneously maximize the utility of all consumers—consumers' objectives are typically in conflict with one another's—so the appropriate notion of optimality is not as straightforward. The notion will we use is that of Pareto optimality, which means that all we are doing is ruling out allocations that are dominated by other feasible allocations.

Definition 3. Given an economy $\mathcal{E}$, a feasible allocation $(x_i)_{i \in \mathcal{I}}$ is **Pareto optimal (or Pareto efficient)** if there is no other feasible allocation $(\hat{x}_i)_{i \in \mathcal{I}}$ such that $u_i(\hat{x}_i) \geq u_i(x_i)$ for all $i \in \mathcal{I}$ with strict inequality for some $i \in \mathcal{I}$.

In words, all Pareto optimality rules out is allocations for which someone could be made better off without making anyone else worse off. This notion of optimality is therefore silent on issues of distribution, since it may be Pareto optimal for one consumer to consume everything in the economy and for everyone else to consume nothing.

Assumptions on Consumer Preferences and Endowments Throughout the next few sections, we will invoke different sets of assumptions for different results. I will collect these assumptions here and will be explicit in referring to them when they are required for a result.

Assumption A1 (continuity): For all consumers $i \in \mathcal{I}$, u_i is continuous.

Assumption A2 (monotonicity): For all consumers $i \in \mathcal{I}$, u_i is increasing: $u_i(x_i') > u_i(x_i)$ whenever $x_{l,i}' > x_{l,ic}$ for all $l \in \mathcal{L}$.

Assumption A3 (concavity): For all consumers $i \in \mathcal{I}$, u_i is concave.

Assumption A4 (interior endowments): For all consumers $i \in \mathcal{I}$, $\omega_{l,i} > 0$ for all $l \in \mathcal{L}$.

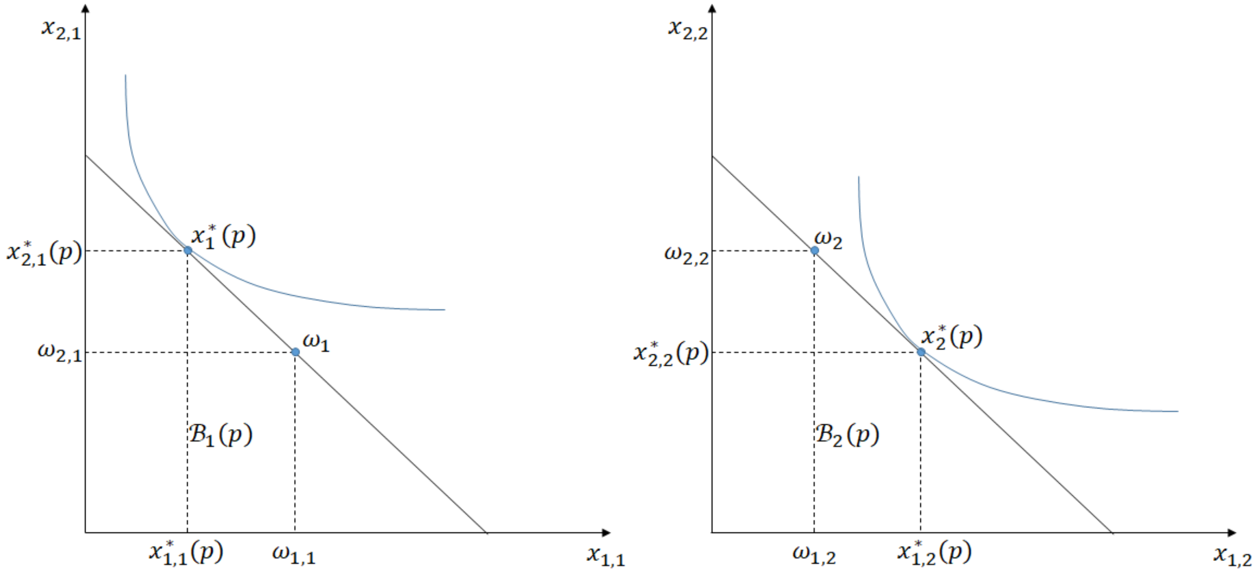
The first three assumptions should be familiar from Economics 2010a. The results we will be establishing in the upcoming sections will hold under weaker assumptions—for example, $(A2)$ can typically be relaxed to local nonsatiation,[1] and $(A3)$ can typically be relaxed to quasiconcavity. The last assumption is a strong assumption that will prove to be sufficient for ruling out some pathological cases in which a Walrasian equilibrium does not exist.

Graphical Examples Many of the main ideas of general equilibrium theory can be understood in a two-consumer, two-commodity pure exchange

[1] We say that consumer i's preferences satisfy **local non-satiation** if for every $x_i \in \mathcal{X}_i$ and every $\varepsilon > 0$, there is an $x_i' \in \mathcal{X}_i$ such that $||x_i' - x_i|| \leq \varepsilon$ and $u_i(x_i') > u_i(x_i)$.

economy. We can get most of the results across graphically in what is referred to as an Edgeworth box. Edgeworth boxes are informationally dense, so let me introduce the constituent elements separately.

Figure 1(a) depicts the relevant information for consumer 1. On the horizontal axis is her consumption of commodity 1 and on the vertical axis is her consumption of commodity 2. Her endowment is $\omega_1 = (\omega_{1,1}, \omega_{2,1})$. At prices $p = (p_1, p_2)$, she can afford to buy any consumption bundle in the set $\mathcal{B}_1(p)$. The slope of her budget line is $-p_1/p_2$. Given her preferences, which are represented by her indifference curve, and given prices p and endowment ω_1, she optimally chooses to consume $x_1^*(p) = \left(x_{1,1}^*(p), x_{2,1}^*(p)\right)$. In other words, at prices p, she would optimally like to sell $\omega_{1,1} - x_{1,1}^*(p)$ units of commodity 1 in exchange for $x_{2,1}^*(p) - \omega_{2,1}$ units of commodity 2. Figure 1(b) depicts the same information for consumer 2.



Figures 1(a) and 1(b): consumer-optimization problems

The Edgeworth box represents both consumers' endowments and their optimal choices as a function of the prices p, so it will incorporate all the information in Figures 1(a) and 1(b). To build towards this goal, Figure 2(a) depicts all the **non-wasteful allocations** in the economy: allocations $(x_i)_{i\in\{1,2\}}$ for which $x_{l,1} + x_{l,2} = \omega_{l,1} + \omega_{l,2}$ for $l \in \{1,2\}$. The bottom-left corner is the origin for consumer 1, and the upper-right corner is the origin for consumer 2. The length of the horizontal axis is equal to the total endowment of commodity 1, and the length of the vertical axis is equal to the total endowment of commodity 2. The horizontal axis, read from the left to the right, represents consumer 1's consumption of commodity 1, and read from the right to the left, represents consumer 2's consumption of commodity 1. The vertical axis, read from the bottom to the top, represents consumer 1's consumption of commodity 2, and read from the top to the bottom, represents consumer 2's consumption of commodity 2. The endowment ω is a point in the Edgeworth box, and it represents a non-wasteful allocation, since $\omega_{l,1} + \omega_{l,2} = \omega_{l,1} + \omega_{l,2}$ for $l \in \{1, 2\}$. The allocation x also represents a non-wasteful allocation, since $x_{l,1} + x_{l,2} = \omega_{l,1} + \omega_{l,2}$ for $l \in \{1, 2\}$.

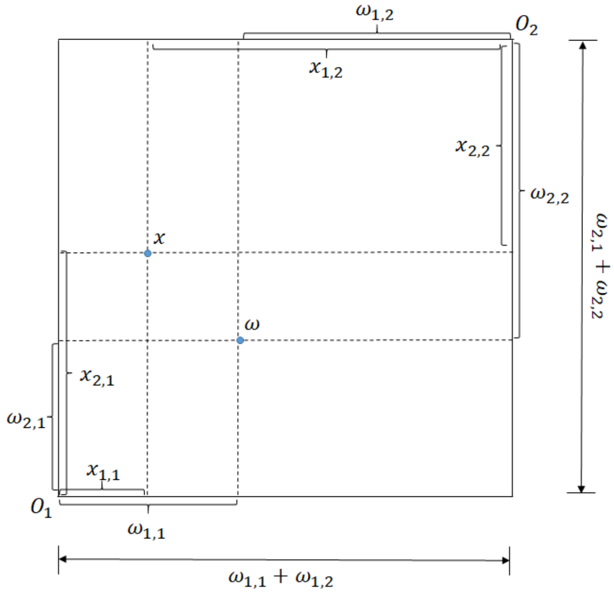


Figure 2(a): non-wasteful allocations

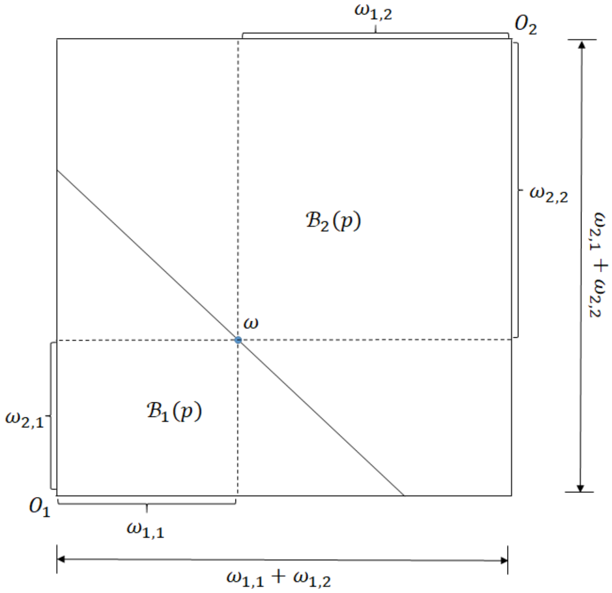
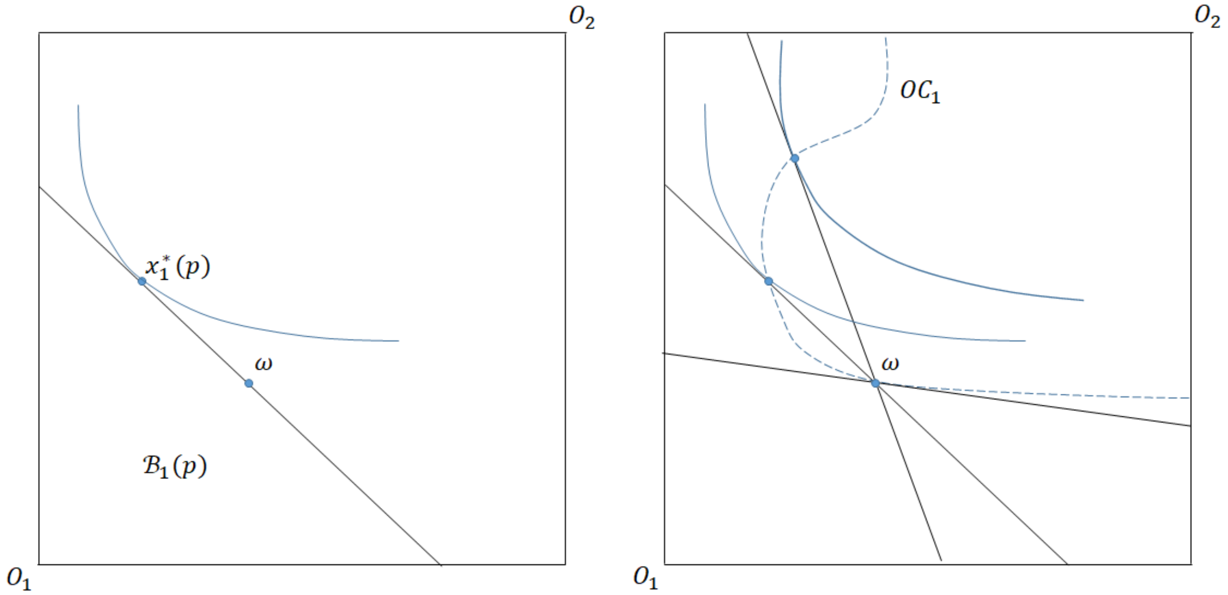


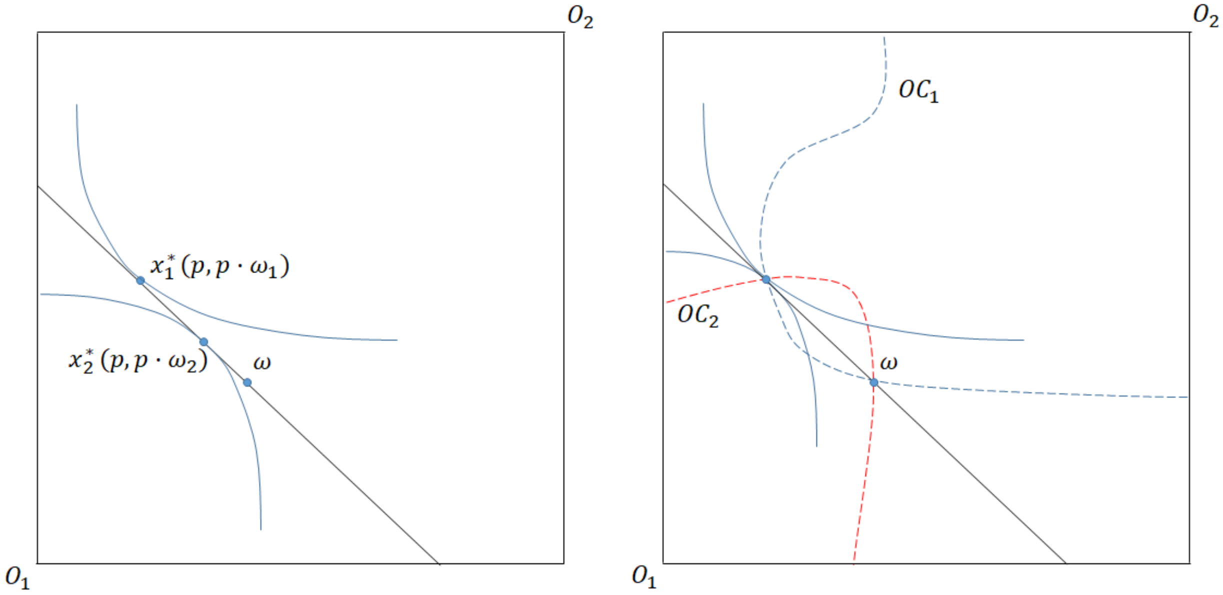
Figure 2(b): prices and budget sets

Figure 2(b) adds prices into the picture and shows that, given any price vector p, the Edgeworth box can be partitioned into the budget sets for the two consumers. Given these prices, consumer 1 can choose any consumption bundle to the bottom left of the diagonal line, and consumer 2 can choose any consumption bundle to the upper right of the diagonal line. Given these prices, Figure 3(a) shows that consumer 1 will optimally choose bundle $x_1^*(p, p \cdot \omega_1)$, and Figure 3(a) shows how $x_1^*(p, p \cdot \omega_1)$ varies as the price ratio varies. Note that, in terms of determining consumer 1's optimal choice, the price ratio p_1/p_2 is a sufficient statistic for the price vector p. This is because Marshallian demand correspondences are homogeneous of degree zero in prices (i.e., $x_1^*(p, p \cdot \omega_1) = x_1^*(\lambda p, \lambda p \cdot \omega_1)$ for all $\lambda \in \mathbb{R}_{++}$).The curve traced out in Figure 3(b) is referred to as consumer 1's **offer curve**.



Figures 3(a) and 3(b): Consumer 1's Marshallian demand for a fixed p and her offer curve

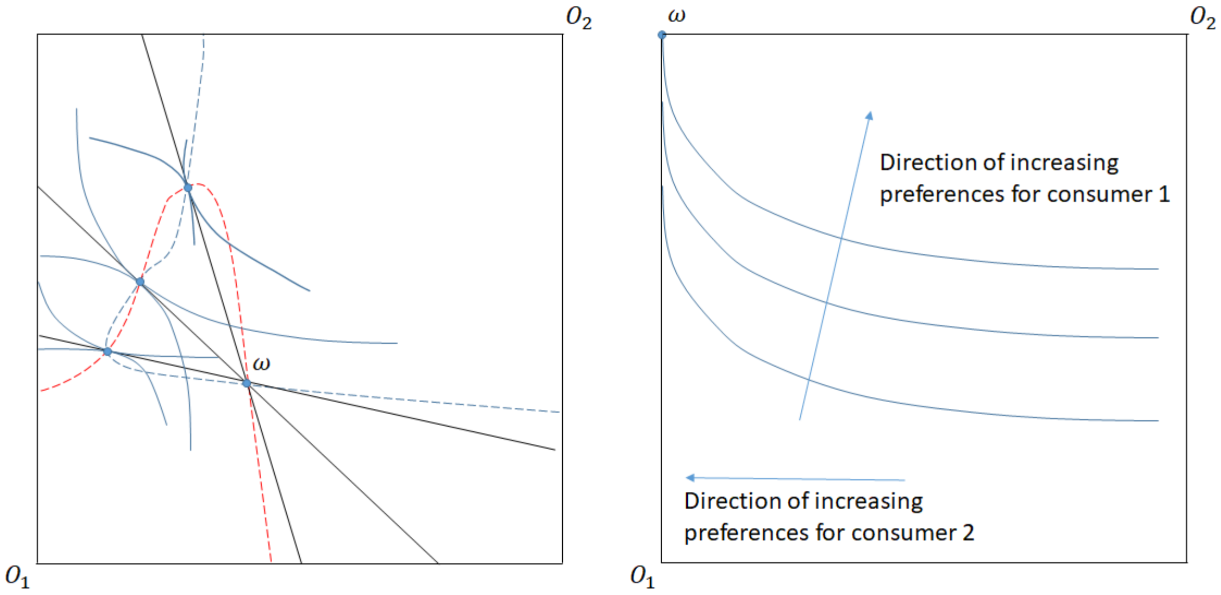
Recall from above that in a Walrasian equilibrium $\left(p^*, (x_i^*)_{i\in\{1,2\}}\right)$, consumer i optimally chooses x_i^* given equilibrium prices p^*. This means that in any Walrasian equilibrium, both consumers' optimal choices lie on their offer curves. Figure 4(a) depicts, for a given price vector p, both consumers' optimal choices. At this price vector, consumer 1 would like to sell $\omega_{1,1} - x_{1,1}^* (p, p \cdot \omega_1)$ units of commodity 1 in exchange for $x_{2,1}^* (p, p \cdot \omega_1) - \omega_{2,1}$ units of commodity 2, and consumer 2 would like to buy $x_{1,2}^* (p, p \cdot \omega_2) - \omega_{1,2}$ units of commodity 1 and sell $\omega_{2,2} - x_{2,2}^* (p, p \cdot \omega_2)$ units of commodity 2. The associated allocation, $(x_1^* (p, p \cdot \omega_1), x_2^* (p, p \cdot \omega_2))$, is not a Walrasian equilibrium allocation, since consumer 1 would like to sell more units of commodity 1 than consumer 2 would like to buy, so the market for commodity 1 does not clear: $\omega_{1,1} - x^*_{1,1}\left(p, p \cdot \omega_1\right) > x^*_{1,2}\left(p, p \cdot \omega_2\right) - \omega_{1,2}$.



Figures 4(a) and 4(b): disequilibrium and equilibrium allocations

It should be clear from the above argument, then, that any Walrasian equilibrium allocation has to occur at a point where both consumers' offer curves intersect. Figure 4(b) illustrates such a point. The upper-left point at which the two offer curves intersect is a Walrasian equilibrium allocation, and the price vector that ensures both players optimally choose the associated consumption bundles is a Walrasian equilibrium price vector. This example also illustrates that, relative to the Walrasian equilibrium allocation, there are no other feasible allocations that can make one of consumers better off without hurting the other consumer. The Walrasian equilibrium allocation is therefore Pareto optimal. Note that the offer curves also intersect at the endowment point, but the endowment point is not a Walrasian equilibrium allocation in this particular example—why?

The Edgeworth box is also useful for illustrating why there may be multiple Walrasian equilibria and why a Walrasian equilibrium might fail to exist. Figure 5(a) illustrates a situation in which there are multiple Walrasian equilibria. Here, the two consumers' offer curves intersect multiple times. Exercise 2 asks you to solve for the set of Walrasian equilibria in another example in which there are multiple equilibria.


O_2
ω
O_1
ω
O_2
Direction of increasing preferences for consumer 1
Direction of increasing preferences for consumer 2
O_1

Figures 5(a) and 5(b): multiple Walrasian equilibria and no Walrasian equilibria

Figure 5(b) illustrates a situation in which there are no Walrasian equilibria. In the example, consumer 1 is endowed with no units of commodity 1 and with all the units of commodity 2. Consumer 2 is endowed with all the units of commodity 1 and with no units of commodity 2. Consumer 2 cares only about her consumption of commodity 1, and consumer 2 cares about both her consumption of commodity 1 and her consumption of commodity 2. Moreover, consumer 1's marginal utility of consuming the first unit of commodity 1 is infinite, and her marginal utility of consuming commodity 2 is strictly positive. For any prices p with $p_1 > 0$, the market for commodity 1 cannot clear, since consumer 2 will always choose $x^*_{1,2}(p, p \cdot \omega_2) = \omega_{1,2}$, and consumer 1 will always choose $x^*_{1,1}(p, p \cdot \omega_1) > 0$, unless $p_2 = 0$. And if $p_2 = 0$, then $x^*_{2,1}(p, p \cdot \omega_1) = +\infty$, so the market for commodity 2 cannot clear. This example illustrates why things can go awry when assumption $(A4)$ is not satisfied. These examples tell us that the answers to the following two important questions is “no”: (a) is there always a Walrasian equilibrium? (b) if there is a Walrasian equilibrium, is it unique?

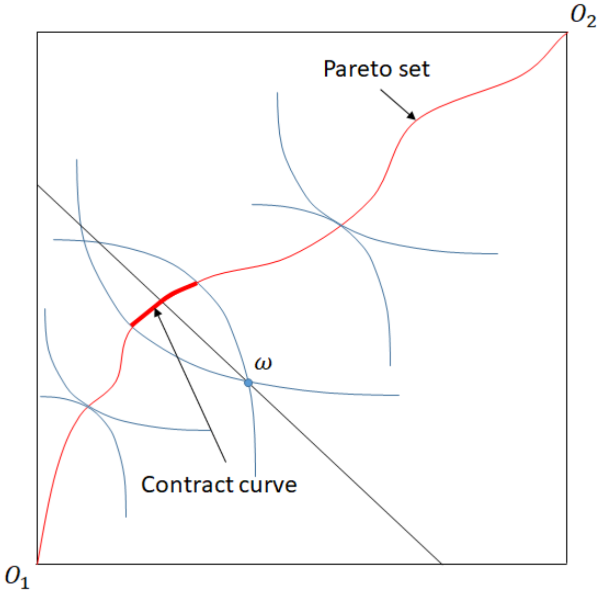


Figure 6: Pareto-optimal allocations and the contract curve

Finally, Figure 6 shows that we can use the Edgeworth box to illustrate the entire set of Pareto-optimal allocations. The **Pareto set** is the set of all feasible allocations for which making one consumer better off necessarily means making the other consumer worse off. It also illustrates the **contract curve**, which is the set of Pareto-optimal allocations that both players prefer to the endowment. If the two consumers were to negotiate a deal, given their endowments as their outside options, they would likely reach a point on the contract curve. Walrasian equilibrium allocations are typically a small subset of the contract curve, and in particular, they lie on the Pareto set. This result is known as the first welfare theorem, and we will now establish this result.

Exercise 1 (Adapted from MWG 15.B.2). Consider an Edgeworth box economy in which the consumers have Cobb-Douglas utility functions $u_1(x_{1,1}, x_{2,1}) = x_{1,1}^{\alpha} x_{2,1}^{1-\alpha}$ and $u_2(x_{1,2}, x_{2,2}) = x_{1,2}^{\beta} x_{2,2}^{1-\beta}$, where $\alpha, \beta \in (0,1)$. Consumer i's endowments are $(\omega_{1,i}, \omega_{2,i}) >> 0$ for $i = 1, 2$. Solve for the Walrasian equilibrium price ratio and allocation. How do these change as you increase $\omega_{1,1}$? Note: Feel free to avoid writing expressions out as much as possible. For example, if you solve for price, feel free to leave the solutions for demand in terms of the price variable instead of plugging in. For comparative statics, if you can find the sign without having to write it out, that's fine.

Exercise 2 (Adapted from MWG 15.B.6). Compute the Walrasian equilibria for the following Edgeworth box economy (there is more than one Walrasian equilibrium):

$$\begin{aligned} u_1(x_{1,1}, x_{2,1}) &= \left(x_{1,1}^{-2} + \left(\frac{12}{37}\right)^3 x_{2,1}^{-2} \right)^{-1/2}, \qquad \omega_1 = (1,0), \\ u_2(x_{1,2}, x_{2,2}) &= \left(\left(\frac{12}{37}\right)^3 x_{1,2}^{-2} + x_{2,2}^{-2} \right)^{-1/2}, \qquad \omega_2 = (0,1). \end{aligned}$$

Exercise 3 (Adapted from MWG 15.B.9). Suppose that in a pure exchange economy, we have two consumers, Alphanse and Betatrix, and two commodities, Perrier and Brie. Alphanse and Betatrix have the utility functions:

$$u_\alpha\left(x_{p,\alpha}, x_{b,\alpha}\right) = \min\left\{x_{p,\alpha}, x_{b,\alpha}\right\} \text{ and } u_\beta\left(x_{p,\beta}, x_{b,\beta}\right) = \min\left\{x_{p,\beta}, \left(x_{b,\beta}\right)^{1/2}\right\},$$

(where $x_{p,\alpha}$ is Alphanse's consumption of Perrier, and so on). Alphanse starts with an endowment of 30 units of Perrier (and none of Brie); Betatrix starts with 20 units of Brie (and none of Perrier). Neither can consume negative amounts of a commodity. If the two consumers behave as price takers, what is the equilibrium? [Hint: consider the market-clearing condition in the cases when both prices are positive, when only the price of Perrier is positive, and when only the price of Brie is positive.]

Exercise 4. Consider an exchange economy with two consumers. The utility functions and endowments are given by

$$\begin{aligned} u_1\left(x_{1,1}, x_{2,1}\right) &= x_{1,1} - \frac{x_{2,1}^{-3}}{3}, \qquad & \omega_1 = (K, r) \\ u_2\left(x_{1,2}, x_{2,2}\right) &= x_{2,2} - \frac{x_{1,2}^{-3}}{3}, \qquad & \omega_2 = (r, K)\,. \end{aligned}$$

Assume that K is sufficiently large so that each consumer can achieve an interior solution to her optimal consumption problem. Note that $p^* = (1, 1)$ is an equilibrium price vector.

(a) For what values of r will there be multiple Walrasian equilibria in this economy? [Hint: first solve for $q = p_y/p_x$ by showing that $rt^4 - t^3 + t - r = 0$, where $t = q^{1/4}$. Note this expression factors as $(t+1)(t-1)(rt^2 - t + r) = 0$.]

(b) For what value of r will $p^* = (1, 3)$ be an equilibrium price vector?

(c) [Optional: algebra intensive] Assume that $K = 10$ and that r takes the value identified in part (b). Find all equilibrium prices and allocations.

(d) [Optional: algebra intensive] Rank the outcomes identified in part (d) in terms of most preferred to least preferred for each consumer.

1.1 First Welfare Theorem

At the Walrasian equilibria in the examples we just saw, there are no feasible allocations that make both players better off: the Walrasian equilibrium allocation was Pareto optimal. This, it turns out, is a general result and perhaps one of the most important results of GE. This result is known as the first welfare theorem. Before stating and proving the first welfare theorem, we will first establish an intermediate result known as Walras's Law, which is a direct implication of consumer optimization when consumers' preferences are monotonic (or more generally, satisfy local non-satiation).

Lemma 1 (Walras's Law). Given an economy $\mathcal{E}$ and prices p, if $(A2)$ holds, then $p \cdot \left(\sum_{i \in \mathcal{I}} x_i \left(p, p \cdot \omega_i\right)\right) = p \cdot \left(\sum_{i \in \mathcal{I}} \omega_i\right)$.

Proof of Lemma 1. Since $(A2)$ holds, each consumer will optimally choose to exhaust her budget: $p \cdot x_i \left(p, p \cdot \omega_i\right) = p \cdot \omega_i$ for all $i \in \mathcal{I}$. Summing this condition over consumers gives us the expression in the Lemma.■

Note that Walras's Law holds for any set of allocations that are consumer-optimal—the result does not require that the allocation $\left(x_i \left(p, p \cdot \omega_i\right)\right)_{i \in \mathcal{I}}$ is a Walrasian equilibrium allocation.

Exercise 5 (Adapted from MWG 15.B.1). Consider an Edgeworth box economy in which the two consumers' preferences satisfy local nonsatiation. Let $x_{l,i} \left(p, p \cdot \omega_i\right)$ be consumer i's demand for commodity l at prices $p = (p_1, p_2)$.

(a) Show that $p_1 \sum_{i \in \mathcal{I}} \left(x_{1,l} - \omega_{1,l}\right) + p_2 \sum_{i \in \mathcal{I}} \left(x_{2,l} - \omega_{2,l}\right) = 0$ for all prices $p \neq 0$.

(b) Argue that if the market for commodity 1 clears at prices $p^* >> 0$, then so does the market for commodity 2; hence p^* is a Walrasian equilibrium price vector.

We can now prove a version of the first welfare theorem.

Theorem 1 (First Welfare Theorem). Suppose $\left(p^*, (x_i^*)_{i\in\mathcal{I}}\right)$ is a Walrasian equilibrium for the economy $\mathcal{E}$. Then if $(A2)$ holds, the allocation $(x_i^*)_{i\in\mathcal{I}}$ is Pareto optimal.

Proof of Theorem 1. In order to get a contradiction, suppose the Walrasian equilibrium allocation $(x_i^*)_{i\in\mathcal{I}}$ is not Pareto optimal. Then there is some other feasible allocation $(\hat{x}_i)_{i\in\mathcal{I}}$ for which $u_i(\hat{x}_i) \geq u_i(x_i^*)$ for all $i \in \mathcal{I}$ and $u_{i'}(\hat{x}_{i'}) > u_{i'}(x_{i'}^*)$ for some i'. Since $(x_i^*)_{i\in\mathcal{I}}$ is a Walrasian equilibrium allocation, and consumers' preferences satisfy $(A2)$, by revealed preference, it has to be the case that $p^* \cdot \hat{x}_i \geq p^* \cdot x_i^*$ for all $i \in \mathcal{I}$ and $p^* \cdot \hat{x}_{i'} > p^* \cdot x_{i'}^*$. Summing up over these conditions,

$$p^* \cdot \left(\sum_{i\in\mathcal{I}} \hat{x}_i\right) > p^* \cdot \left(\sum_{i\in\mathcal{I}} x_i^*\right) = p^* \cdot \left(\sum_{i\in I} \omega_i\right),$$

where the equality holds by Lemma 1. Since equilibrium prices p^* are nonnegative (why are they nonnegative?), this inequality implies that there is some commodity l such that $\sum_{i\in\mathcal{I}} \hat{x}_{i,l} > \sum_{i\in\mathcal{I}} \omega_{i,l}$, and therefore $(\hat{x}_i)_{i\in\mathcal{I}}$ is not a feasible allocation.■

The first welfare theorem is a remarkable result because (a) its conclusion is both intellectually important and powerful, (b) its explicit assumptions are quite weak, and (c) it has a simple proof, in the sense that it involves only a couple steps, and each step is completely transparent. Let me comment a bit more on each of these three points.

The first welfare theorem provides a formal statement of a version of Adam Smith's argument that the "invisible hand" of decentralized markets leads selfish consumers to make decisions that lead to socially efficient outcomes. Despite there being no explicit coordination among consumers, the resulting equilibrium allocation is Pareto optimal.

Second, the only explicit assumption we made in order to prove the first welfare theorem was that consumers have monotonic preferences—and even this assumption can be relaxed, as exercise 6 below asks you to show. But in the background, there are several strong and important assumptions. First, we assumed that all consumers face the same prices as each other for all commodities. Second, we assumed that all consumers are price takers—they take prices as given and understand that their consumption decisions do not affect these prices. Third, there are markets for each commodity, and all consumers can freely participate in each market. Fourth, we assumed that each consumer cares only about her own consumption and not about the consumption of anyone else in the economy—we have therefore ruled out externalities. Finally, we assumed that there are a finite number of commodities and consumers. Exercise 6 asks you to show that when there are an infinite number of commodities and consumers, Walrasian equilibrium allocations need not be Pareto optimal.

Exercise 6 (Adapted from MWG 16.C.3). In this exercise, you are asked to establish the first welfare theorem under a set of assumptions compatible with satiation. First, we will define the appropriate notion of equilibrium. Given an economy $\mathcal{E}$, an allocation $(x_i^*)_{i\in\mathcal{I}}$ and a price vector $p = (p_1, \dots, p_L)$ constitutes a **price equilibrium with transfers** if there is an assignment of wealth levels $(w_1, \dots, w_I)$ with $\sum_{i\in\mathcal{I}} w_i = p \cdot \left(\sum_{i\in\mathcal{I}} \omega_i\right)$ such that: (i) consumers optimize: $x_i^*(p, w_i) = x_i^*$ and (ii) markets clear: $\sum_{i\in\mathcal{I}} x_i^* = \sum_{i\in\mathcal{I}} \omega_i$. Suppose that every $\mathcal{X}_i$ is nonempty and convex and that every u_i is strictly convex. Prove the following:

(a) For every consumer i, there is at most one consumption bundle at which she is locally satiated. Such a bundle, if it exists, uniquely maximizes u_i on $\mathcal{X}_i$.

(b) Any price equilibrium with transfers is a Pareto optimum.

Exercise 7. This exercise illustrates that the importance of the assumption that there are a finite number of commodities for the first welfare theorem. Consider an economy in which there is one physical good, available at infinitely many dates: $t = 1, 2, \dots,$ so there are effectively an infinite number of commodities: the physical good at date 1, the physical good at date 2, and so on. One consumer (or "generation") is born at each date $t = 1, 2, \dots,$ and lives and consumes at dates t and $t + 1$ ("young" and "old"). We will refer to the consumer born in date t as consumer t. There is also one old consumer alive at date $t = 1$ (call her consumer 0), and she is endowed with zero units of the good. Each consumer is endowed with one unit of the good when she is young, no units of the good when she is old, and no storage is possible. Consumption in each period is non-negative, and each consumer t's preferences over consumption is given by $u_t(x_{t,t}, x_{t+1,t}) = u(x_{t,t}) + u(x_{t+1,t})$, where u is smooth, increasing, and strictly concave, with $u'(0) < \infty$.

(a) Show that there is a Walrasian equilibrium in which each consumer consumes her endowment and gets utility $u(1) + u(0)$.

(b) Show that the above Walrasian equilibrium is unique.

(c) Show that the above Walrasian equilibrium allocation is not Pareto optimal. In other words, construct a feasible allocation that is strictly better for each consumer.

I want to conclude this section with a couple comments on the simplicity of the proof of the first welfare theorem. First, the theorem itself presents a partial characterization of equilibrium allocations. To prove the statement, we did not need to solve explicitly for a Walrasian equilibrium and show that it is Pareto optimal. Instead, we described properties that all Walrasian equilibrium allocations must satisfy. Second, the statement itself is a conditional statement. It is a statement of the form "if $(x_i^*)_{i\in\mathcal{I}}$ is a Walrasian equilibrium, then $(x_i^*)_{i\in\mathcal{I}}$ is Pareto optimal." This conditional statement dodges the question of whether there is in fact a Walrasian equilibrium—we showed above that there does not always exist a Walrasian equilibrium, and we will spend some time next week providing conditions under which a Walrasian equilibrium in fact exists. Finally, the proof is a proof by contradiction, and it effectively takes the form of "if this Walrasian equilibrium allocation was not Pareto-optimal, then stuff doesn't add up." While elegant, the proof itself provides little insight into *why* the first welfare theorem holds. We will spend a little more time discussing the "why" in week 3.

1.2 Second Welfare Theorem

The first welfare theorem establishes that Walrasian equilibrium allocations are Pareto optimal. The second welfare theorem in some sense establishes a converse. It says that, under some assumptions, any Pareto optimal allocation can be "decentralized" as a Walrasian equilibrium allocation, given the correct prices and endowments.

Theorem 2 (Second Welfare Theorem). Let $\mathcal{E}$ be an economy that satisfies $(A1) - (A4)$ and $\mathcal{X}_i = \mathbb{R}^L_+$. If $(\omega_i)_{i\in\mathcal{I}}$ is Pareto optimal, then there exists a price vector $p \in \mathbb{R}^L_+$ such that $\left(p, (\omega_i)_{i\in\mathcal{I}}\right)$ is a Walrasian equilibrium for $\mathcal{E}$.

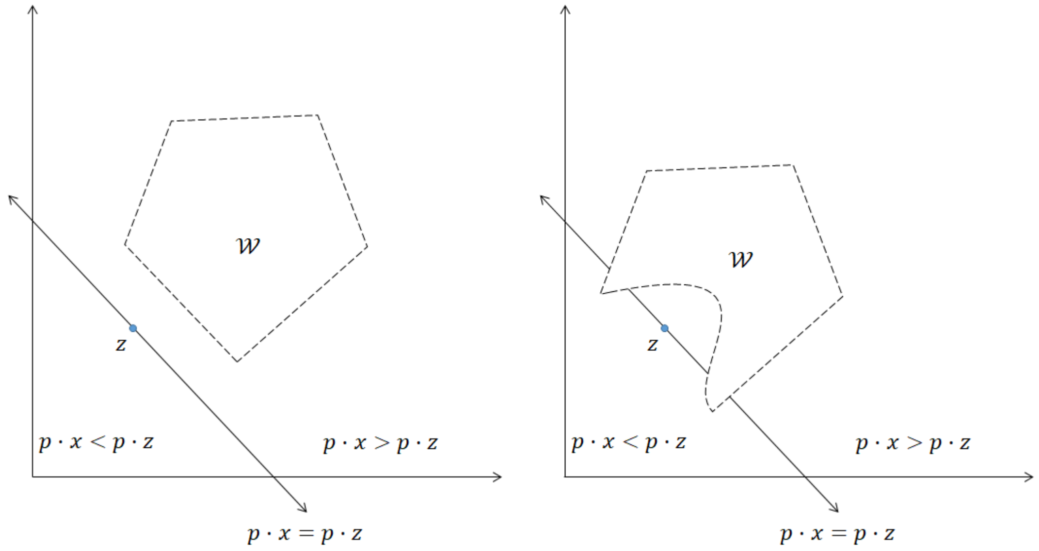
Before proving the second welfare theorem, we will state a version of an important theorem in convex analysis, which is used in the key step of the proof of the second welfare theorem.

Lemma 2 (Separating Hyperplane Theorem). If $\mathcal{W} \subseteq \mathbb{R}^n$ is an open convex set, and $z \notin \mathcal{W}$ is a point not in $\mathcal{W}$, then there exists a vector $p \neq 0$ and such that $p \cdot x \geq p \cdot z$ for all $x \in cl\,(\mathcal{W})$.

Figure 7(a) illustrates this version of the separating hyperplane theorem in two dimensions. The set $\mathcal{W}$ is open and convex, and $z \notin \mathcal{W}$. The point z is on a line (which is a hyperplane in a two-dimensional space) characterized by the equation $p \cdot x = p \cdot z$ (i.e., p is the normal vector to the line). All the points to the upper right of that line satisfy $p \cdot x > p \cdot z$, and all the points to the lower left of that line satisfy $p \cdot x < p \cdot z$. And in particular, $\mathcal{W}$ is fully to the upper right of this line. In the case illustrated in Figure 7(a), there are of course many other separating hyperplanes satisfying $p \cdot x \geq p \cdot z$ for all $x \in \mathcal{W}$ corresponding to differently sloped lines going through z but not intersecting $\mathcal{W}$. Figure 7(b) shows why the assumption that $\mathcal{W}$ is a convex set is important for this result. If $z \notin \mathcal{W}$, but $z \in conv\,(\mathcal{W})$,[2] then there is

[2] The set $conv\,(\mathcal{W})$ is defined to be the smallest convex set containing $\mathcal{W}$. In two dimensions, you can visualize $conv\,(\mathcal{W})$ by taking $\mathcal{W}$ and putting a rubber band around it.

no vector $p \neq 0$ for which $p \cdot x \geq p \cdot z$ for all $x \in \mathcal{W}$. Exercise 8 asks you to prove a stronger version of the separating hyperplane theorem, which shows that any two disjoint convex sets can be separated by a hyperplane.



Figures 7(a) and 7(b): separating hyperplane theorem and convexity.

The idea of the second welfare theorem is to show that, if the endowment $(\omega_i)_{i \in \mathcal{I}}$ is Pareto optimal, we can always find a price vector that separates the set of allocations preferred by all consumers in the economy from $(\omega_i)_{i \in \mathcal{I}}$ and therefore show that $\left(p, (\omega_i)_{i \in \mathcal{I}}\right)$ is a Walrasian equilibrium.

Proof of Theorem 2. By the statement of the theorem, $(\omega_i)_{i \in \mathcal{I}}$ is Pareto optimal. Let us define the set of aggregate consumption bundles that can be allocated in such a way among consumers to make them all strictly better off than under $(\omega_i)_{i\in\mathcal{I}}$. To do so, define the set of consumption bundles that consumer i prefers to ω_i:

$$\mathcal{A}_i = \left\{a \in \mathbb{R}^L : a + \omega_i \geq 0 \text{ and } u_i\left(a + \omega_i\right) > u_i\left(\omega_i\right)\right\}.$$

Since u_i is concave, the set $\mathcal{A}_i$ is convex. The Minkowski sum of the sets $\mathcal{A}_i$ is therefore also a convex set.[3] That is, if we define

$$\mathcal{A} = \sum_{i\in\mathcal{I}} \mathcal{A}_i = \left\{a \in \mathbb{R}^L : \exists a_1 \in \mathcal{A}_1, \ldots, \exists a_I \in \mathcal{A}_I \text{ with } a = \sum_{i\in\mathcal{I}} a_i\right\},$$

then $\mathcal{A}$ is a convex set. The set $\mathcal{A}$ does not contain the 0 vector because $(\omega_i)_{i\in\mathcal{I}}$ is Pareto optimal. To see why this is the case, note that if $0 \in \mathcal{A}$, then there would exist $(a_i)_{i\in\mathcal{I}}$ with $\sum_{i\in\mathcal{I}} a_i = 0$ and $u_i\left(a_i + \omega_i\right) > u_i\left(\omega_i\right)$ for all i. That is, we could essentially just reallocate the endowment $(\omega_i)_{i\in\mathcal{I}}$ among the I consumers and make them all strictly better off, but that would contradict the assumption that $(\omega_i)_{i\in\mathcal{I}}$ is Pareto optimal.

Next, by Lemma 2, there is some price vector $p^* \neq 0$ such that $p^* \cdot a \geq 0$ for all $a \in cl\left(\mathcal{A}\right)$. Moreover, each of the prices $p_l^* \geq 0$. To see why, suppose $p_l^* < 0$ for some l. Take some a for which a_l is arbitrarily large and all other $a_{l'}$ are arbitrarily small but positive. By the monotonicity of consumer

[3] The Minkowski sum of two sets $\mathcal{A}$ and $\mathcal{B}$ is just the set of vectors x that can be written as the sum of vectors $x = a + b$ for which $a \in \mathcal{A}$ and $b \in \mathcal{B}$. The closest visual analog to thinking about the Minkowski sum of sets in two dimensions is the way the clone stamp tool in Photoshop works if you are familiar with it.

preferences, $a \in \mathcal{A}$, but if a is chosen this way, then $p \cdot a < 0$. We therefore have that $p^* > 0$ (i.e., $p_l^* \geq 0$ for all $l \in \mathcal{L}$ with at least one inequality strict).

We will now show that $\left(p^*, (\omega_i)_{i \in \mathcal{I}}\right)$ is a Walrasian equilibrium. To do so, we need to show that at p^*, consumers optimally consume their endowments and that markets clear. The second condition is immediate. It remains to show that at this p^*, consumers optimally consume their endowments. To do so, suppose there is some $\hat{x}_i \in \mathbb{R}_+^L$ for which $u_i(\hat{x}_i) > u_i(\omega_i)$. We will show that this $\hat{x}_i \notin \mathcal{B}_i(p^*)$. By the definition of $\mathcal{A}$, the allocation $(x_i)_{i \in \mathcal{I}} - (\omega_i)_{i \in \mathcal{I}}$ with $x_i = \hat{x}_i$ and $x_j = \omega_j$ for all $j \neq i$, is in $cl(\mathcal{A})$. By the definition of p^*, we necessarily have that $p^* \cdot (\hat{x}_i - \omega_i) + p^* \cdot \sum_{j \neq i} (\omega_j - \omega_j) \geq 0$, which implies that $p^* \cdot \hat{x}_i \geq p^* \cdot \omega_i > 0$, where this last inequality holds because of Assumption $(A4)$ that all consumers have positive endowments of all commodities.

We are not yet done, because we have to show that this last inequality is strict. This is where continuity of preferences (Assumption $(A1)$) comes into the picture. Since $u_i(\hat{x}_i) > u_i(\omega_i)$, this implies that for λ just less than 1, $u_i(\lambda \hat{x}_i) > u_i(\omega_i)$, which in turn implies that $\lambda p^* \cdot \hat{x}_i \geq p^* \cdot \omega_i > 0$. This cannot be the case if $p^* \cdot \hat{x}_i = p^* \cdot \omega_i$, so we must therefore have that $p^* \cdot \hat{x}_i > p^* \cdot \omega_i$ and hence $\hat{x}_i \notin \mathcal{B}_i(p^*)$—that is, any allocation preferred by consumer i to her endowment is unaffordable, and hence her optimal consumption bundle is her endowment.■

The second welfare theorem does not show that every Pareto optimal allocation is a Walrasian equilibrium given a *particular* endowment. Instead, it says that if we were to start from a particular endowment $(\omega_i)_{i \in \mathcal{I}}$, and an allocation $(x_i)_{i\in\mathcal{I}}$ is Pareto optimal, then we could reallocate consumers' endowments in such a way that $(x_i)_{i\in\mathcal{I}}$ is a Walrasian equilibrium allocation. The version of the theorem that we just proved carries out this exercise using a particularly stark reallocation of endowments (i.e., it just sets $(\omega_i)_{i\in\mathcal{I}} = (x_i)_{i\in\mathcal{I}}$). There are versions of the theorem that involve carrying out lump-sum transfers of wealth rather than directly moving around endowments. As you might expect, decentralizing a particular Pareto-optimal allocation in practice potentially requires large-scale redistribution of wealth. I view the result more as establishing an equivalence between Walrasian equilibria and Pareto-optimal allocations rather than as a practical guide for figuring out how to achieve a particular distribution of consumption in society.

It is worth a reminder that convexity of consumers' preferences was critical in establishing the result that $\mathcal{A}$ was a convex set, which in turn is required for using the separating hyperplane theorem. Figure 8 shows an example where the conclusion of the second welfare theorem fails if consumers' preferences are not convex.

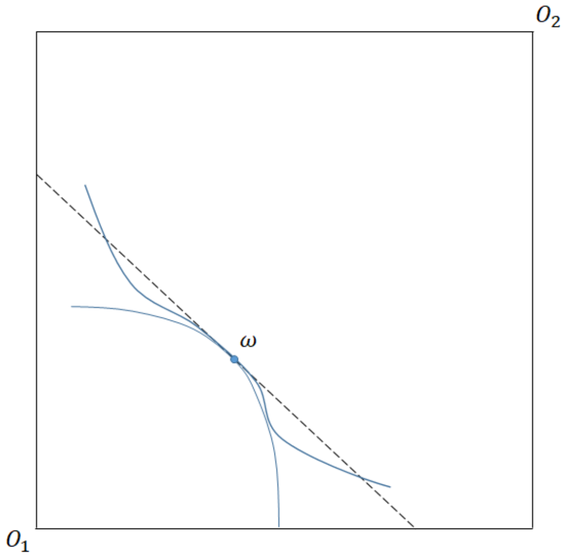


Figure 8: Non-convex preferences

In this figure, the endowment is a Pareto-optimal allocation, since consumer 1's and consumer 2's better-than sets are separated. But there are no prices that can make it optimal for consumer 1 to consume ω_1.

Nevertheless, a version of the second welfare theorem continues to hold when consumers do not have convex preferences if you replicate the economy a large number of times. Think of the 2-consumer economy as being a metaphor for a large economy with two *types* of consumers: type-1 consumers have preferences u_1 and endowments ω_1, and type-2 consumers have preferences u_2 and endowments ω_2. If we replicate the economy a large number of times, so that there are N type-1 consumers and N type-2 consumers, where N is large, then we can support ω as a Walrasian equilibrium allocation, at least on average. This result follows from an application of the Shapley-Folkman lemma, which roughly says that the Minkowski average of sets converges to the convex hull of that set. You don't need to know the math behind this result, but it is a useful result to be aware of. Figure 9 illustrates a replication economy for the economy described in Figure 8. It shows that there may be a p for which there is an $x_1 \in x_1(p, p \cdot \omega_1)$ and an $x_1' \in x_1(p, p \cdot \omega_1)$, so that if we allocate a fraction λ of type-1 consumers to consume x_1 and a fraction $1 - \lambda$ of type-1 consumers to consume x_1', on average they are consuming ω_1: $\lambda x_1 + (1 - \lambda) x_1' = \omega_1$.

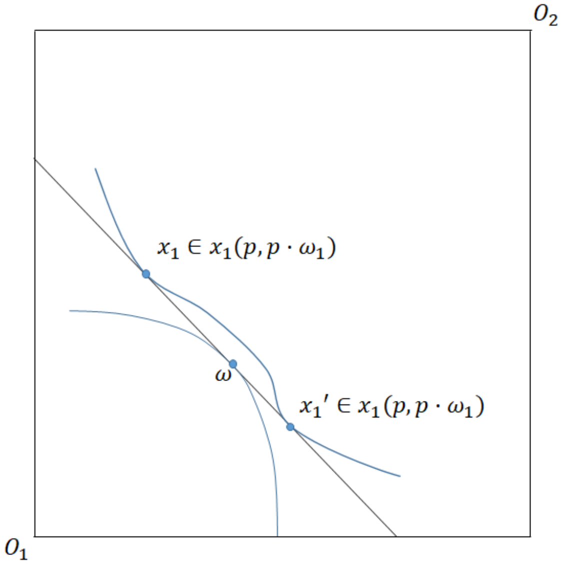

O_2
$x_1 \in x_1(p, p \cdot \omega_1)$
ω
$x_1{}' \in x_1(p, p \cdot \omega_1)$
O_1

Figure 9: Replication economy

This figure illustrates the idea that large numbers "convexifies" the economy. There is a recurring theme throughout general equilibrium theory that many of the pathologies that arise seem to "go away" in sufficiently large economies. Nonconvexities seem esoteric, since we usually think of consumers' preferences as having diminishing marginal utility and preferences for variety. Nonconvexities become especially relevant when we think of firms, though. When there are fixed costs, for example, the firm analogue of consumers' "better-than" sets are not convex, since the set of production levels better than "not even breaking even" can include both "shut down" and "produce, but at a much larger scale."

Finally, we made use of Assumption $(A4)$ in a somewhat opaque way in the proof. What Assumption $(A4)$ rules out is cases like the one illustrated in Figure 5(b) in which there were no Walrasian equilibria. The failure of equilibrium existence illustrated in Figure 5(b) arises because of a sort of "division by zero" problem: supporting the endowment as an equilibrium allocation would have required consumer 2 to buy only a finite amount of a commodity with a zero price when she has zero wealth.

Exercise 8. This question is intended to guide you through a proof of the separating hyperplane theorem. This is more of an exercise in math than in economics, so feel free to skip to the next step if you get stuck.

(a) Prove that if $y \in \mathbb{R}^N$ and $\mathcal{C} \subseteq \mathbb{R}^N$ is closed, then there exists a point $z \in \mathcal{C}$ such that $||z - y|| \leq ||x - y||$ for all $x \in \mathcal{C}$. That is, there exists a point in $\mathcal{C}$ that is closest to y. (You may assume that $||\cdot||$ is the Euclidean norm.) Hint: use the Weierstrass extreme value theorem—if f is a real-valued and continuous function on domain $\mathcal{S}$, and $\mathcal{S}$ is compact and non-empty, then there exists x such that $f(x) \geq f(y)$ for all $y \in \mathcal{S}$.

(b) Suppose further that $\mathcal{C} \subseteq \mathbb{R}^N$ is convex, and note from above that if $y \notin \mathcal{C}$, then there exists $z \in \mathcal{C}$ that is closest to y. Let $x \in \mathcal{C}$ with $x \neq z$.

(i) Show that $(y - z) \cdot z \geq (y - z) \cdot x$. Hint: consider $||y - (z + t(x - z))||$ for $t \in [0, 1]$, the distance between y and a convex combination of x and z.

(ii) Use the above result to show that for all $x \in \mathcal{C}$, $(y - z) \cdot y > (y - z) \cdot x$.

(iii) Explain how this is a special case of the separating hyperplane theorem, which states that for any disjoint convex sets $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^N$, there exists nonzero $p \in \mathbb{R}^N$ such that $p \cdot u \geq p \cdot v$ for any $u \in \mathcal{A}$ and $v \in \mathcal{B}$.

(iv) Use the result of (ii) to deduce the separating hyperplane theorem. Hint: consider $y = 0$ and $\mathcal{C} = \mathcal{A} - \mathcal{B} = \{u - v : u \in \mathcal{A}, v \in \mathcal{B}\}$.

1.3 Characterizing Pareto-Optimal Allocations

The welfare theorems provide a tight connection between the set of Pareto optimal allocations and the set of Walrasian equilibrium allocations. This section will provide a short note on how to find Pareto optimal allocations in particularly well-behaved environments. Define the **utility possibility set**

$$\mathcal{U} = \left\{ (u_1, \ldots, u_I) \in \mathbb{R}^I : \text{there is a feasible allocation } (x_i)_{i \in \mathcal{I}} \text{ with } u_i(x_i) \geq u_i \text{ for all } i \right\}.$$

If the sets $\mathcal{X}_i$ are convex sets and consumers' preferences are concave, then $\mathcal{U}$ is a convex set. When this is the case, the problem of finding Pareto-optimal allocations can be reduced to the problem of solving **Pareto problems** of the form

$$\max_{u \in U} \lambda \cdot u$$

for some non-zero vector of **Pareto weights** $\lambda \geq 0$. The objective function of this problem is sometimes called a linear Bergson-Samuelson social welfare function. We will say that u^* is a **Pareto-optimal utility vector** if there is a Pareto-optimal allocation $(x_i)_{i \in \mathcal{I}}$ for which $u_i(x_i) = u_i^*$ for all $i \in \mathcal{I}$. The next theorem establishes the result.

Theorem 3. If u^* is a solution to the Pareto problem described above for some vector of Pareto weights $\lambda >> 0$, then u^* is a Pareto-optimal utility vector. Conversely, if the utility possibility set $\mathcal{U}$ is convex, then any Pareto-optimal utility vector u^* is a solution to the Pareto problem for some non-zero vector $\lambda \geq 0$.

Proof of Theorem 3. The first part is immediate: if u^* is not Pareto optimal, then any Pareto-dominating utility vector would give a higher value in the Pareto problem for any Pareto weight vector $\lambda >> 0$.

The second part of the theorem makes use of the *supporting hyperplane theorem*, which says that a convex set can be separated from any point outside its interior (see Section M.G of the mathematical appendix of MWG). If u^* is a Pareto-optimal utility vector, then it lies on the boundary of $\mathcal{U}$, so by the supporting hyperplane theorem, there exists $\lambda \neq 0$ such that $\lambda \cdot u^* \geq \lambda \cdot u$ for all $u \in \mathcal{U}$. Further, the Pareto weights satisfy $\lambda \geq 0$, since if $\lambda_i < 0$ for some i, then $\lambda \cdot u^* < \lambda \cdot \tilde{u}$, where for some $K > 0$, $\tilde{u} = \left(u_1^*, \ldots, u_{i-1}^*, u_i^* - K, u_{i+1}^*, \ldots, u_I^*\right) \in \mathcal{U}$. This contradicts the claim that $\lambda \cdot u^* \geq \lambda \cdot u$ for all $u \in \mathcal{U}$, so it must be the case that $\lambda \geq 0$.■

The theorem shows that when the utility possibility set is a convex set, the problem of finding Pareto-optimal allocations boils down to solving a class of Pareto problems. If we further assume that consumers' utility functions are differentiable, then Pareto-optimal allocations can be characterized by taking first-order conditions. For example, suppose utility functions are differentiable with $\nabla u_i(x_i) >> 0$ for all x_i, and we have an interior solution, we can find Pareto-optimal allocations by solving the problem:

$$\max_{(x_i)_{i \in \mathcal{I}}} \sum_{i \in \mathcal{I}} \lambda_i u_i(x_i)$$

subject to feasibility for each commodity:

$$\sum_{i \in \mathcal{I}} x_{l,i} \leq \sum_{i \in \mathcal{I}} \omega_{l,i} \text{ for all } l \in \mathcal{L}.$$

Then one can use the Kuhn-Tucker theorem to verify that any Pareto-optimal allocation $(x_i)_{i \in \mathcal{I}}$ with $x_i >> 0$ for all $i \in \mathcal{I}$ must satisfy

$$\frac{\partial u_i / \partial x_{l,i}}{\partial u_i / \partial x_{l',i}} = \frac{\partial u_{i'} / \partial x_{l,i'}}{\partial u_{i'} / \partial x_{l',i'}} = \frac{\mu_l}{\mu_{l'}} \text{ for all } i, i', l, l'$$

for some $\mu_l, \mu_{l'} > 0$. This condition says that the marginal rate of substitution between any two commodities must be equalized across consumers in any Pareto-optimal allocation. If this condition failed, there would be a Pareto-improving exchange of commodities l and l' between consumers i and i'. The values μ_l corresponds to the Lagrange multiplier on the commodity-l

feasibility constraint $\sum_{i\in\mathcal{I}} x_{l,i} = \sum_{i\in\mathcal{I}} \omega_{l,i}$.

As an illustration of the second welfare theorem, given a Pareto-optimal allocation $(x_i)_{i\in\mathcal{I}}$ that satisfies the optimality conditions above, if you set $p_l = \mu_l$ for all $l \in \mathcal{L}$, then $\left(p, (x_i)_{i\in\mathcal{I}}\right)$ is a Walrasian equilibrium of the economy $\mathcal{E} = \left((u_i)_{i\in\mathcal{I}}, (x_i)_{i\in\mathcal{I}}\right)$. This point is illustrated in Figure 4(b). In that figure, at the Walrasian equilibrium allocation, consumers' marginal rates of substitution across the two commodities were equalized. Moreover, these marginal rates of substitution were also equal to the price ratio that corresponded to the Walrasian equilibrium (and given that price ratio, moving the endowment along the boundary of consumers' budget sets does not change their ultimate consumption choices, so the same price vector would also be an equilibrium price vector if we just set consumers' endowments equal to their Walrasian equilibrium allocations).

Exercise 9 (Adapted from MWG 16.C.4). Suppose that for each consumer, there is a "happiness function" depending on her own consumption only, given by $u(x_i)$. Every consumer's utility depends positively on her own and everyone else's "happiness" according to the utility function

$$U_i(x_1, \ldots, x_l) = U_i(u_1(x_1), \ldots, u_l(x_l)).$$

Show that if $x = (x_1, \ldots, x_l)$ is Pareto optimal relative to the $U_i(\cdot)$'s, then $x = (x_1, \ldots, x_l)$ is also a Pareto optimum relative to the u_i's. Does this mean a community of altruists can use competitive markets to attain Pareto optima? Does your argument depend on the concavity of the u_i's or the U_i's?

1.4 Existence of Walrasian Equilibrium

Last week, we focused on the normative, efficiency properties of Walrasian equilibria. This week, we will focus on a couple positive properties. In particular, we will begin by asking what seems like a straightforward question: does a Walrasian equilibrium exist? And then we will ask a few other important questions relating to equilibrium uniqueness, equilibrium stability, and the comparative statics of Walrasian equilibria.

The question of whether a Walrasian equilibrium exists really boils down to: under what conditions on preferences and endowments does a Walrasian equilibrium exist? We know from the example in Figure 5(b) from last week that a Walrasian equilibrium does not always exist. And we know from the second welfare theorem that when assumptions $(A1) - (A4)$ are satisfied, then if the endowment is a Pareto-optimal allocation, there is a Walrasian equilibrium for which it is the equilibrium allocation. In some sense, the second welfare theorem provides a bit of a mundane answer to the existence question, since it provides conditions under which no trade is optimal for each consumer. The more interesting question is the more difficult one: when is a Walrasian equilibrium guaranteed to exist if the endowment itself is not already Pareto optimal? That is, when is there a Walrasian equilibrium that actually involves trade?

This was an open question ever since Walras's formulation of the GE model in the 1870's until Arrow, Debreu, and McKenzie produced the first rigorous existence proofs in the 1950's. The basic question is, given aggregate demand functions $\sum_{i\in\mathcal{I}} x_{l,i}\left(p, p\cdot\omega_i\right)$ for each commodity, when does there exist a price vector p^* such that $\sum_{i\in\mathcal{I}} x_{l,i}\left(p^*, p^*\cdot\omega_i\right) = \sum_{i\in\mathcal{I}} \omega_{l,i}$ for all $l \in \mathcal{L}$? Early arguments amounted to just counting up the number of equations and unknowns, but these approaches were not satisfactory, since it would not be clear what would happen if the solution to the equations involved negative prices or quantities. The breakthrough came in the 1950's when Arrow and Debreu (1954) proved the following existence result.

Theorem 4 (Existence of Walrasian Equilibrium). Given an economy $\mathcal{E}$ satisfying $(A1)-(A4)$, there exists a Walrasian equilibrium $\left(p^*, \left(x_i^*\right)_{i\in\mathcal{I}}\right)$.

The key insight in the 1950's was to reframe the Walrasian equilibrium existence question as a *fixed-point* question, following John Nash's (1951) proof of the existence of Nash equilibrium using a related approach. A *fixed point* of a correspondence $f : Z \rightrightarrows Z$ is a point z such that $z \in f\left(z\right)$, and fixed-point theorems provide fairly general conditions under which functions or correspondences have fixed points.

The important step in making use of the general-purpose technology of fixed-point theorems is to figure out how to map the equilibrium existence question into the question of whether a suitably chosen correspondence has a fixed point: it is about choosing the right correspondence. Suppose the correspondence f maps an allocation $\left(x_i\right)_{i\in\mathcal{I}}$ and a price vector p into a new allocation $\left(x_i'\right)_{i\in\mathcal{I}}$ and price vector p', where the new allocation is the set of optimal choices for consumers given the price vector p, and the new price vector

p' is one that raises the prices of over-demanded goods and lowers the price of under-demanded goods under the allocation $(x_i)_{i\in\mathcal{I}}$ and otherwise does not change prices. Then a fixed point of f will be a Walrasian equilibrium, so if we can show that f satisfies the conditions required for a fixed-point theorem to apply, then we can conclude that a Walrasian equilibrium exists.

Two-Commodity Intuitive Sketch

We will first go through an intuitive argument for equilibrium existence in the special case of two commodities, and then we will go through the more general result described in Theorem 4. The argument in the two-commodity case will also develop some tools that will be useful when we talk about uniqueness and stability of Walrasian equilibrium. For this part, we will strengthen the monotonicity condition $(A2)$ to a strong monotonicity condition $(A2')$.

Assumption A2' (strong monotonicity). For all consumers $i \in \mathcal{I}$, u_i is strictly increasing: $u_i(x_i') > u_i(x_i)$ whenever $x_{l,i}' \geq x_{l,i}$ for all $l \in \mathcal{L}$ with at least one inequality strict.

As a starting point, we are going to introduce the idea of an excess demand function for an economy $\mathcal{E} = (u_i, \omega_i)_{i\in\mathcal{I}}$. The **excess demand function for consumer** i is $z_i(p) = x_i(p, p\cdot\omega_i) - \omega_i$. The **aggregate excess demand function** is the sum of consumers' excess demand functions $z(p) = \sum_{i\in\mathcal{I}} z_i(p)$. It should be clear from the definition of the aggregate excess demand function that if there is a p^* that satisfies $z(p^*) = 0$, then

$\left(p^*, (x_i^*)_{i\in\mathcal{I}}\right)$ with $x_i^* = x_i(p^*, p^* \cdot \omega_i)$ is a Walrasian equilibrium. The way x_i^* is defined, it is clear that at p^*, x_i^* is consumer i's optimal consumption bundle. Moreover, if $z(p^*) = 0$, then markets clear for each commodity. In this case, proving existence boils down to establishing that a solution to $z(p) = 0$ exists given assumptions $(A1), (A2'), (A3)$, and $(A4)$. The aggregate excess demand function inherits many of the properties of Marshallian demand functions, as the next lemma illustrates.

Lemma 3. Suppose $\mathcal{E}$ satisfies $(A1), (A2'), (A3)$,$(A4)$, and $\mathcal{X}_i = \mathbb{R}_+^L$ for all i. Then the aggregate excess demand function satisfies:

(i) z is continuous;

(ii) z is homogeneous of degree zero;

(iii) $p \cdot z(p) = 0$ for all p (Walras's Law);

(iv) there is some $Z > 0$ such that $z_l(p) > -Z$ for every $l \in \mathcal{L}$ and for every p;

(v) if $p^n \to p$, where $p \neq 0$ and $p_l = 0$ for some l, then $\max\{z_1(p^n), \ldots, z_L(p^n)\} \to \infty$.

The first property is something that was assumed in Economics 2010a, but it is straightforward to show that it follows from assumptions $(A1) - (A4)$.[4] The second property is straightforward, and we already proved the third property in week 1. The fourth property follows directly from the assumption that $\mathcal{X}_i = \mathbb{R}_+^L$.

[4] z is upper hemi-continuous from Berge's maximum theorem, it is non-empty because preferences are continuous, and it is convex-valued because preferences are convex. These properties imply that z is a continuous correspondence.

The last property bears some comment. It is saying that as some, but not all, prices go to zero, there must be some consumer whose wealth is not going to zero. Because she has strongly monotone preferences, she must demand more of one of the commodities whose price is going to zero.

To gain intuition for the general existence proof, let us consider the case where there are only two goods in the economy, and let us further assume that consumer preferences are strictly concave, so that $x_i\left(p, p \cdot \omega_i\right)$ is a singleton for all p (we will allow for $x_i\left(p, p \cdot \omega_i\right)$ to be a correspondence—for there to be multiple optimal allocations for a given consumer at a given price vector—when we prove the general theorem). Our goal is to find a price vector $p = (p_1, p_2)$ for which $z\left(p\right) = 0$. Because $z\left(\cdot\right)$ is homogeneous of degree zero, we can normalize one of the prices, say p_2, to one. This reduces our search to price vectors of the form $(p_1, 1)$. Moreover, Walras's Law implies that if the market for commodity 1 clears, then so does the market for commodity 2, so it suffices to find a price p_1 such that $z_1\left(p_1, 1\right) = 0$.

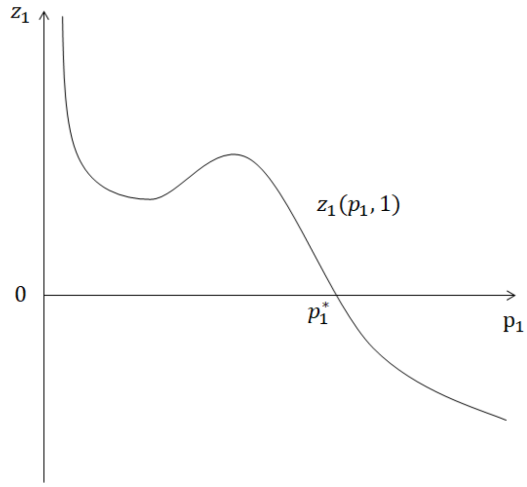


Figure 10: Existence of WE with two commodities

The problem of finding a p_1 such that $z_1(p_1, 1) = 0$ is a one-dimensional problem, so we can just graph it. Figure 10 plots $z_1(p_1, 1)$ as a function of p_1. The figure highlights three important properties of $z_1(p_1, 1)$. First, it is continuous. Second, for p_1 very small, $z_1(p_1, 1) > 0$, and third, for p_1 very large, $z_1(p_1, 1) < 0$. Given these three properties, by the intermediate value theorem—the simplest of fixed-point theorems—there necessarily exists some $p^* = (p_1^*, 1)$ such that $z_1(p^*) = 0$, and a Walrasian equilibrium therefore exists. The subtleties in making this argument are in establishing that $z_1(p_1, 1) > 0$ for p_1 small and $z_1(p_1, 1) < 0$ for p_1 large. The first property follows from condition (v) in the Lemma above. The second property follows because if $p_1 \to \infty$, then each consumer's demand for commodity

1 will converge to something less than her endowment of commodity 1, as continuity and monotonicity of preferences imply she would like to sell at least some of commodity 1 for an unboundedly large amount of commodity 2.

Exercise 10 (Adapted from MWG 17.C.4). Consider a pure exchange economy. The only novelty is that a progressive tax system is instituted according to the following rule: individual wealth is no longer $p \cdot \omega_i$; instead, anyone with wealth above the mean of the population must contribute half of the excess over the mean into a fund, and those below the mean receive a contribution from the fund in proportion to their deficiency below the mean.

(a) For a two-consumer society with endowments $\omega_1 = (1, 2)$ and $\omega_2 = (2, 1)$, write the after-tax wealths of the two consumers as a function of prices.

(b) If the consumer preferences are continuous, strictly convex, and strongly monotone, will the excess demand functions satisfy the conditions required for existence stated in Lemma 3?

More General Existence Result

Before proving the main existence theorem, we will first remind ourselves of a couple important mathematical theorems that we will be using in the proof. The first is the Kakutani fixed-point theorem, which you used to prove the existence of a Nash equilibrium in Economics 2010a. The second is the maximum theorem or Berge's maximum theorem.

Kakutani Fixed-Point Theorem. Suppose $\mathcal{Z}$ is a nonempty, compact, convex subset of $\mathbb{R}^n$ and that $f : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is a nonempty, convex-valued, and upper hemi-continuous correspondence. Then f has a fixed point.

Kakutani's fixed-point theorem is a generalization of Brouwer's fixed-point theorem but for set-valued functions. The basic idea of the theorem is that a fixed point is an intersection of the graph of f with the 45° line, and the conditions for the theorem ensure that the graph of f cannot "jump" across the 45° line. In the special case when $n = 1$ and when f is scalar-valued, this theorem boils down to the intermediate-value theorem.

The proof of equilibrium existence is going to make use of the Kakutani fixed-point theorem for an appropriately defined correspondence, and we will need to be able to establish that the correspondence has the properties that are required by the theorem. The following theorem will be useful for establishing these properties.

Berge's Maximum Theorem. If $f : \mathcal{X} \times \Theta \to \mathbb{R}$ is a continuous function, and $C : \Theta \rightrightarrows \mathcal{X}$ is a continuous, compact-valued correspondence, then $V(\theta) = \max\{f(x, \theta) : x \in C(\theta)\}$ is continuous in θ, and $X^*(\theta) = \operatorname{argmax}\{f(x, \theta) : x \in C(\theta)\}$ is non-empty, compact-valued, and upper hemi-continuous.

The line of proof we will be following is to define a set $\mathcal{Z}$ and a correspondence $f : \mathcal{Z} \rightrightarrows \mathcal{Z}$ that satisfies the conditions of Kakutani fixed-point theorem and whose fixed points are Walrasian equilibria. There are therefore three main questions we will need to answer:

1. What should $\mathcal{Z}$ and f be?

2. Why do the conditions of Kakutani's fixed-point theorem hold?

3. Why do the fixed points of f correspond to Walrasian equilibria?

Step 1: Define $\mathcal{Z}$ and f To define the set $\mathcal{Z}$, it is convenient to first normalize prices so that they sum to one. Define the normalized price simplex Δ to be the set of associated price vectors: $\Delta \equiv \left\{p \in \mathbb{R}_{+}^{L} : \sum_{l \in \mathcal{L}} p_l = 1\right\}$. Next, for each consumer i, define a non-empty, compact, convex subset of her consumption set that is bounded above by what she could consume if she possessed the entire aggregate endowment: $\mathcal{T}_i = \left\{x_i \in \mathcal{X}_i : x_i \leq 2 \sum_{i \in \mathcal{I}} \omega_i\right\} \subset \mathcal{X}_i$. Since each $\mathcal{T}_i$ is a compact, so is the product set $\mathcal{T} = \prod_{i \in \mathcal{I}} \mathcal{T}_i$. Define $\mathcal{Z} \equiv \mathcal{T} \times \Delta$ to be the domain on which we will define the correspondence f.

The correspondence $f : \mathcal{Z} \rightrightarrows \mathcal{Z}$ will map an allocation $(x_i)_{i \in \mathcal{I}}$ and a price vector p to the set of allocations $(x_i')_{i \in \mathcal{I}}$ that are optimal for each consumer given p and a new price vector p' that raises the price of commodities that were over-demanded and lowers the price of commodities that were under-demanded under $(x_i)_{i \in \mathcal{I}}$. The first part of this construction is straightforward. Let $x_i^{\mathcal{T}}(p, p \cdot \omega_i)$ be consumer i's optimal choice over $\mathcal{B}_i(p) \cap \mathcal{T}_i$. Marshallian demand correspondence at prices p:

$$x_i^{\mathcal{T}}(p, p \cdot \omega_i) = \max_{x_i \in \mathcal{B}_i(p) \cap \mathcal{T}_i} u_i(x_i),$$

where recall that $\mathcal{B}_i(p)$ is consumer i's budget set given prices p:

$$\mathcal{B}_i(p) = \{x_i \in \mathcal{X}_i : p \cdot x_i \leq p \cdot \omega_i\}.$$

If we do this for each consumer, we get the product of correspondences $\prod_{i\in\mathcal{I}} x_i^{\mathcal{T}}\left(p, p\cdot\omega_i\right) \subset \mathcal{T}$. This takes care of the first part of the construction.

For the second part of the construction, we introduce a fictitious "player" called the Walrasian auctioneer (or the "price player") who chooses a price vector $p \in \mathbb{R}_+^L$ and wants to maximize the value of aggregate excess demand. Let

$$a^*\left(x\right) = \underset{\tilde{p}\in\Delta}{\operatorname{argmax}}\, \tilde{p}\cdot\left(\sum_{i\in\mathcal{I}} x_i - \sum_{i\in\mathcal{I}}\omega_i\right),$$

where "a^*" is a pneumonic for "auctioneer." We are now in a position to define the appropriate correspondence $f:\mathcal{Z}\rightrightarrows\mathcal{Z}$ by

$$f\left(x,p\right) = \underbrace{\left(\prod_{i\in\mathcal{I}} x_i^{\mathcal{T}}\left(p, p\cdot\omega_i\right)\right)}_{\subset\mathcal{T}} \times \underbrace{a^*\left(x\right)}_{\subset\Delta}.$$

Step 2: Verify that Kakutani's theorem can be applied Now that we have defined the set $\mathcal{Z}$ and the correspondence f, we will verify that the conditions of Kakutani's fixed-point theorem hold, so it can be applied. The first set of conditions that needs to be verified is that $\mathcal{Z} = \mathcal{T}\times\Delta$ is a non-empty, compact, convex subset of $\mathbb{R}^n$ for some n. The second set of conditions is on the correspondence f—we need to show that f is a non-empty, convex-valued, and upper hemi-continuous correspondence. Note that the product of non-empty, convex-valued, and upper hemi-continuous correspondences is itself non-empty, convex-valued, and upper hemi-continuous, so this last part requires that we show that each of the correspondences $x_i^{\mathcal{T}}(p, p \cdot \omega_i)$ and $a^*(x)$ satisfy these conditions.

First, note that $\mathcal{Z}$ is a non-empty, compact, and convex subset of $\mathbb{R}^{I \cdot L + L}$ because each $\mathcal{T}_i$ and Δ are non-empty, compact, and convex subsets of $\mathbb{R}^L$.

Next, a^* is non-empty and convex-valued because Δ is non-empty, compact, and convex, and the Walrasian auctioneer's objective is linear in p and hence continuous. It is upper hemi-continuous by Berge's maximum theorem.

Finally, the function $x_i^{\mathcal{T}}$ is non-empty and convex-valued because $\mathcal{B}_i(p) \cap \mathcal{T}_i$ is non-empty, compact, and convex, and u_i is continuous (guaranteeing x_i^* is non-empty) and concave (guaranteeing x_i^* is convex-valued). These conditions alone are not enough to give us the upper hemi-continuity that we require in order to apply Kakutani's fixed-point theorem, however, because we still have to show that $\mathcal{B}_i(p) \cap \mathcal{T}_i$ is a continuous, compact-valued correspondence.

It is apparent that $\mathcal{B}_i(p) \cap \mathcal{T}_i$ is compact-valued—the involved part is showing that it is a continuous correspondence. To do so, we have to show that it is both upper hemi-continuous in p and lower hemi-continuous in p. Upper hemi-continuity is straightforward, since if $p^n \to p$ and $x_i^n \to x_i$ with $x_i^n \in \mathcal{B}_i(p^n) \cap \mathcal{T}_i$ for all n, then $p^n \cdot x_i^n \leq p^n \cdot \omega_i$ and $x_i^n \leq \sum_{i \in \mathcal{I}} \omega_i$ for all n and therefore this condition holds in the limit as well. Showing that $\mathcal{B}_i(p) \cap \mathcal{T}_i$ is lower hemi-continuous is more involved, and we leave this as an exercise (this is the only part of the existence proof that makes use of assumption $(A4)$).

These arguments establish that $\mathcal{Z}$ and f satisfy the conditions of Kakutani's fixed-point theorem, and therefore f has a fixed point. We now need to show that any such fixed point of f is a Walrasian equilibrium.

Step 3: Show that fixed points of f are Walrasian equilibria Suppose $(x^*, p^*) \in f(x^*, p^*)$—that is, (x^*, p^*) is a fixed point of f. We need to show that $x_i^* = x_i^{\mathcal{T}}(p^*, p^* \cdot \omega_i)$ is consumer-optimal for each $i \in \mathcal{I}$, and markets clear at prices p^*.

For the first part, because $x_i^{\mathcal{T}}(p^*, p^* \cdot \omega_i) = \max_{x_i \in \mathcal{B}_i(p^*) \cap \mathcal{T}_i} u_i(x_i)$, we need to verify that the resulting solution also solves the relaxed problem $\max_{x_i \in \mathcal{B}_i(p^*)} u_i(x_i)$, which is the problem the consumer actually faces. To do this, first note that since consumers have monotonic preferences, it must be the case that $p^* \cdot \left(\sum_{i \in \mathcal{I}} x_i^*\right) \leq p^* \cdot \left(\sum_{i \in \mathcal{I}} \omega_i\right)$—if we did not have to worry about $x_i \in \mathcal{T}_i$ for each i, this inequality would hold with equality by Walras's Law. Next, since $p^* \in a^*(x^*)$, we have

$$0 \geq p^* \cdot \left(\sum_{i \in \mathcal{I}} x_i^* - \sum_{i \in \mathcal{I}} \omega_i\right) \geq p \cdot \left(\sum_{i \in \mathcal{I}} x_i^* - \sum_{i \in \mathcal{I}} \omega_i\right) \text{ for all } p \in \Delta,$$

so that $\sum_{i \in \mathcal{I}} x_i^* - \sum_{i \in \mathcal{I}} \omega_i \leq 0$ and therefore $x_i^* \leq \sum_{i \in \mathcal{I}} \omega_i$ for all i, so that $x_i^* \in int(\mathcal{T}_i)$. We therefore have that $x_i^* \in \operatorname{argmax}_{x_i \in \mathcal{B}_i(p^*)} u_i(x_i)$ because if there were some $\hat{x}_i \in \mathcal{B}_i(p^*)$ with $u_i(\hat{x}_i) > u_i(x_i^*)$, then for some small λ, $\lambda \hat{x}_i + (1 - \lambda) x_i^* \in \mathcal{B}_i(p^*) \cap \mathcal{T}_i$ and by (quasi-)concavity of u_i, $u_i(\lambda \hat{x}_i + (1 - \lambda) x_i^*) > u_i(x_i^*)$, which is a contradiction.

We now establish that the market-clearing condition is satisfied. Since x_i^* is consumer-optimal for each i, Walras's Law tells us that $p^* \cdot \left(\sum_{i \in \mathcal{I}} x_i^*\right) = p^* \cdot \left(\sum_{i \in \mathcal{I}} \omega_i\right)$, and in particular, at p^*, the Walrasian auctioneer's value is zero (recall that the auctioneer maximizes $p \cdot \left(\sum_{i \in \mathcal{I}} x_i - \sum_{i \in \mathcal{I}} \omega_i\right)$). If $\sum_{i \in \mathcal{I}} x_{l,i}^* - \sum_{i \in \mathcal{I}} \omega_{l,i}$ were positive for any commodity l, then the auctioneer could set $p_l = 1$ and $p_{l'} = 0$ for all $l' \neq l$ and attain a positive value. This implies that no commodity is over-demanded at the allocation x^*, that is, $\sum_{i \in \mathcal{I}} x_i^* \leq \sum_{i \in \mathcal{I}} \omega_i$.

It remains only to show that this inequality actually holds with equality. By Walras's law, we know that $p^* \cdot \left(\sum_{i \in \mathcal{I}} x_i^*\right) = p^* \left(\sum_{i \in \mathcal{I}} \omega_i\right)$. Since there is no excess demand, this implies that commodity l can be in excess supply only if its price is $p_l^* = 0$. In that case, we can just modify the allocation x^* by giving the entire excess supply of commodity l to some consumer—without loss of generality, let that be consumer 1. This is feasible, and it does not affect consumer 1's utility. Why doesn't it affect her utility? Since her preferences are monotone, giving her more of commodity l cannot decrease her utility. It also cannot increase her utility, because otherwise, she would have chosen the resulting consumption bundle rather than x_i^*, and doing so would have been affordable, because $p_l^* = 0$.

To summarize, either (x^*, p^*) is a Walrasian equilibrium or the allocation resulting from arbitrarily allocating any commodity in excess supply to consumers (along with the price vector p^*) is a Walrasian equilibrium. In either case, a Walrasian equilibrium exists.■

1.5 Uniqueness, Stability, and Testability

We now provide an introduction to some of the most important positive properties of general equilibrium theory. We will ask when a Walrasian equilibrium is unique, whether it is stable in the sense that it can be reached by a simple price adjustment process, and we will look at whether Walrasian equilibrium imposes substantive restrictions on observable data.

This lecture will be less formal than previous lectures, mostly going through each of these topics at a rather high level. We have already alluded to the answers to some of these questions: no, Walrasian equilibria need not be unique, and no, it is not the case that a simple price adjustment process will always converge to a Walrasian equilibrium. We will first establish these results under general preferences. We will then focus on a special class of economies in which consumer preferences satisfy the *gross substitutes property*—when this property is satisfied, the model is particularly well-behaved: there will be a unique Walrasian equilibrium, and there will be a simple price-adjustment process that will always converge to it.

Uniqueness and Stability under Fairly General Preferences

Uniqueness We will first look at the question of whether there is a globally unique Walrasian equilibrium. Recall from the previous lecture the definition of the aggregate excess demand function $z(p) = \sum_{i \in \mathcal{I}} z_i(p)$, where $z_i(p) =$

$x_i(p, p \cdot \omega_i) - \omega_i$.

Let us consider a two-commodity, two-consumer economy and normalize $p_2 = 1$. We argued informally last time that a Walrasian equilibrium exists by claiming that $z_1(p_1, 1)$ is continuous in p_1, $z_1(p_1, 1) > 0$ for p_1 small, and $z_1(p_1, 1) < 0$ for p_1 large. By the intermediate-value theorem, there exists a p_1^* such that $z_1(p_1^*, 1) = 0$ and therefore $(p_1^*, 1)$ is a Walrasian equilibrium price vector.

Is there any reason to think that there is only one p_1^* at which $z_1(p_1^*, 1) = 0$? Yes, if $z_1(p_1, 1)$ is everywhere downward-sloping, and in some sense, this is the natural case. It just says that there is less aggregate demand for commodity 1 when p_1 is higher, and we will show later that when the economy satisfies the gross substitutes condition, this will always be the case. But there certainly are situations where $z_1(p_1, 1)$ is not always downard-sloping. There is the somewhat pathological case in which commodity 1 is a Giffen good, so that $x_{1,i}(p, w)$ is increasing in p_1 even holding w fixed. Even if neither good is a Giffen good, however, $x_{1,i}(p, p \cdot \omega_1)$ may be increasing in p_1 because consumer i's wealth is increasing in p_1, so an upward-sloping region of $z_1(p_1, 1)$ is not particularly implausible.

In the first lecture, we discussed an example in the Edgeworth box in which the two consumers' offer curves intersected at three equilibrium points, and in the first problem set, you were asked to solve for the set of Walrasian equilibria in a numerical example for which there were three equilibria. Recall the example from the problem set. Consumers' preferences and endowments are:

$$\begin{aligned} u_1\left(x_{1,1},x_{2,1}\right) &= \left(x_{1,1}^{-2}+\left(\frac{12}{37}\right)^3 x_{2,1}^{-2}\right)^{-1/2}, \qquad \omega_1=(1,0)\,,\\ u_2\left(x_{1,2},x_{2,2}\right) &= \left(\left(\frac{12}{37}\right)^3 x_{1,2}^{-2}+x_{2,2}^{-2}\right)^{-1/2}, \qquad \omega_2=(0,1)\,. \end{aligned}$$

If we normalize $p_2 = 1$, consumers' Marshallian demands for commodity 1 are:

$$x_{1,1}\left(\left(p_1,1\right),p_1\right)=\frac{p_1}{p_1+\frac{12}{37}p_1^{1/3}},\ x_{1,2}\left(\left(p_1,1\right),1\right)=\frac{1}{p_1+\frac{37}{12}p_1^{1/3}},$$

and the aggregate excess demand for commodity 1 is therefore

$$z_1\left(p_1,1\right)=\frac{1}{p_1+\frac{37}{12}p_1^{1/3}}-\frac{\frac{12}{37}p_1^{1/3}}{p_1+\frac{12}{37}p_1^{1/3}}.$$

Figure 11 plots $z_1\left(p_1,1\right)$ and shows that there are three solutions to $z_1\left(p_1,1\right)=0$. You might recall from the first problem set that for $p_1^*\in\left\{\frac{27}{64},1,\frac{64}{27}\right\}$, there is a Walrasian equilibrium with prices $(p_1^*,1)$.

There are two additional general points that you can see illustrated in both Figures 5(a) and 11. The first is that if you were to perturb the economy slightly by changing consumers' preferences or endowments by a tiny amount, this would not affect the fact that there are three Walrasian equilibria.

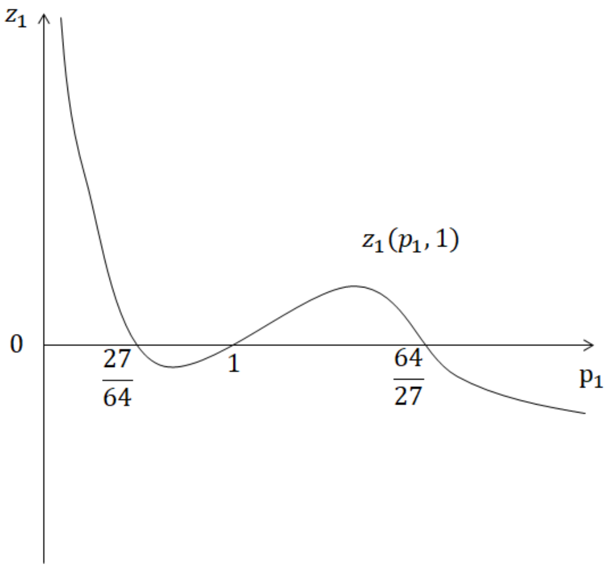


Figure 11: Multiple Walrasian equilibria

The second general point that these examples illustrate is that even though Walrasian equilibria may not be globally unique, they may be what is referred to as *locally unique* in the sense that there is no other Walrasian equilibrium price vector within a small enough range around the original equilibrium price vector. Figure 12 illustrates an example for which this is not the case. An equilibrium is not locally unique if its price vector p is the limit of a sequence of other equilibrium price vectors. This example shows that this can happen, but only if $z_1(p_1, 1)$ is flat and equal to zero over some interval of prices $[p_1^*, p_1^{**}]$. The important point to note about this example is that it is not *generic*: any small perturbation of $z_1(\cdot, 1)$ that would arise from, say, a change in endowments, would restore the property that there are a finite number of equilibria.

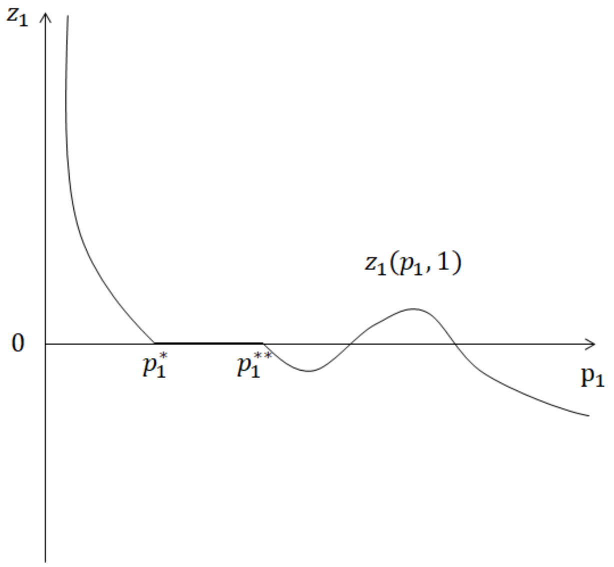


Figure 12: Walrasian Equilibria need not be locally unique

Exercise 11 (Adapted from MWG 17.D.1). Consider an exchange economy with two commodities and two consumers. Both consumers have homothetic preferences of the constant elasticity variety. Moreover, the elasticity of substitution is the same for both consumers and is small (i.e., commodities are close to perfect complements). Specifically,

$$u_1\left(x_{1,1},x_{2,1}\right)=\left(2x_{1,1}^{\rho}+x_{2,1}^{\rho}\right)^{1/\rho} \text{ and } u_2\left(x_{1,2},x_{2,2}\right)=\left(x_{1,2}^{\rho}+2x_{2,2}^{\rho}\right)^{1/\rho},$$

and $\rho=-4$. The endowments are $\omega_1=(1,0)$ and $\omega_2=(0,1)$. Compute the excess demand function of this economy and find the set of competitive equilibria.

Tatonnement Stability One important aspect of Walrasian equilibrium that we have alluded to throughout the course but have not yet addressed is: where do Walrasian equilibrium prices come from? General equilibrium theory is quite weak on the kinds of price-adjustment processes that might lead to Walrasian equilibrium outcomes.

Walras proposed a process he called "tatonnement" whereby a fictitious Walrasian auctioneer gradually raises the price of commodities in excess demand and reduces the prices of those in excess supply until markets clear. This process is related to what the Walrasian auctioneer did in our proof of existence from last time, but not quite the same. In particular, the process last time adjusted prices discontinuously, but it was aimed at showing the existence of a fixed point for a particular operator, not at showing that the fixed point(s) of that operator could be found by iterating it from an arbitrary starting point.

Formally, consider the following continuous-time price-adjustment process $p(t)$:

$$\frac{dp(t)}{dt} = \alpha z(p(t)),$$

for some constant $\alpha > 0$. Given a starting price vector $p(0)$, the process raises prices for any commodities l for which $z_l(p(t)) > 0$ (i.e., for which there is excess demand), and it reduces prices for those for which $z_l(p(t)) < 0$.

The stationary points of this process are prices p at which $z(p) = 0$: Walrasian equilibrium prices. An equilibrium price vector p^* is said to be *locally* *stable* if the price-adjustment process converges to p^* from any "nearby" price vectors, and it is *globally stable* if the process converges to p^* from *any* initial starting price vector. Does this process converge to a Walrasian equilibrium price vector? When there are only two commodities, and the economy satisfies properties $(A1)$, $(A2')$, $(A3)$, and $(A4)$, this process does in fact converge, as Figure 13 highlights. Here, we can also see that p_1^* and p_1^{***} are locally stable, and p_1^{**} is not, and none of the equilibrium price vectors is globally stable.

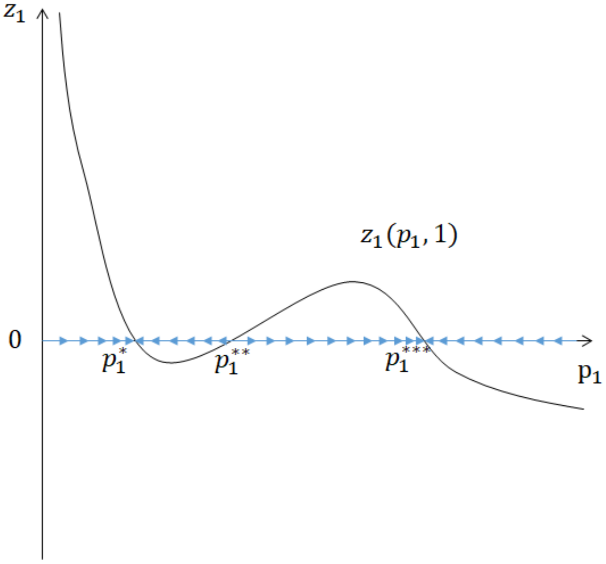


Figure 13: Tatonnement process for two commodities

This price-adjustment process gives us a way to study how equilibrium prices might be reached, but it has several drawbacks. First, the process itself is a conceptual exercise rather than a practical one—the GE model predicts that no one will trade at non-equilibrium prices. Second, if one were to try to implement this process by asking consumers how much they would demand at different price levels, then they would be unlikely to want to report their demands truthfully. Finally, the main drawback with this procedure is that it does not in general converge to an equilibrium price vector. In a famous paper, Scarf (1960) provided several examples in which the process does not converge when there are more than two commodities. We will show in the next section, however, that there are classes of economies for which it does converge.

Uniqueness and Stability under Gross Substitutes

In this section, we will show that economies that satisfy the gross substitutes property have particularly nice properties: there is a unique Walrasian equilibrium (up to a normalization), it is globally stable, and it has nice comparative statics properties.

Recall from consumer theory that commodities k and l are gross substitutes if an increase in p_k increases the Marshallian demand for commodity l (and vice versa), holding wealth fixed. The analogous definition in general equilibrium is as follows.

Definition 4. A Marshallian demand function $x\left(p, p \cdot \omega\right)$ satisfies **gross**

substitutes at endowment ω if, for all prices p and p' with $p'_k > p_k$ and $p'_l = p_l$ for all $l \neq k$, we have $x_l\left(p', p' \cdot \omega\right) > x_l\left(p, p \cdot \omega\right)$ for all $l \neq k$.

This definition of gross substitutes is more subtle than the definition you saw from consumer theory, since increasing p_k also increases the consumer's wealth. It is straightforward to show that if all commodities are gross substitutes in the consumer-theory sense and they are also normal goods (so that demand increases with wealth), then demand functions will satisfy the gross substitutes property for all possible (non-negative) endowments. It is not readily apparent from the definition of gross substitutes that the demand for commodity l is decreasing in p_l, but it is true: since demand is homogeneous of degree 0 in p, increasing p_l is the same as holding p_l fixed and decreasing all other prices. Since decreasing each of these other prices decreases demand for commodity l, so does decreasing all of them.

If each consumer i's demand function satisfies gross substitutes at ω_i, then so does aggregate demand $\sum_{i \in \mathcal{I}} x_i\left(p, p \cdot \omega_i\right)$. The property is restrictive, but it is satisfied by many common functional forms such as CES preferences: $u_i\left(x_i\right) = \left(\sum_{l \in \mathcal{L}} \alpha_l x_{l,i}^{\rho}\right)^{1/\rho}$ for $0 < \rho < 1$.

If aggregate demand satisfies the gross substitutes property, then there is a unique Walrasian equilibrium, as the following result shows.

Proposition 1. If the aggregate excess demand function $z\left(\cdot\right)$ satisfies gross substitutes, the economy has at most one Walrasian equilibrium (up to a normalization).

Proof of Proposition 1. We need to show that there is at most one (normalized) price vector p such that $z(p) = 0$. To see why this is the case, suppose $z(p) = z(p') = 0$ for two price vectors p and p' that are not collinear. By homogeneity of degree zero, we can normalize the price vectors in such a way that $p'_l \geq p_l$ for all $l \in \mathcal{L}$ and $p'_k = p_k$ for some commodity k. Then, to move from p to p', we can think about doing this in $L-1$ steps, increasing the prices of each commodity $l \neq k$ in turn. At each step where a component of the price vector increases strictly, the aggregate demand for commodity k must strictly increase, so that $z_k(p') > z_k(p) = 0$. Moreover, there must be at least one such k, since p is not collinear with p', yielding a contradiction.■

When aggregate demand satisfies the gross substitutes property, not only is there a unique Walrasian equilibrium, but the tatonnement price-adjustment process we described above globally converges to it. To establish this result, we will first prove a lemma.

Lemma 4. Suppose that the aggregate excess demand function $z(\cdot)$ satisfies gross substitutes and that $z(p^*) = 0$. Then for any p not collinear with p^*, $p^* \cdot z(p) > 0$.

Proof of Lemma 4. We will give the proof in the $L = 2$ case. Normalize $p_2 = p_2^* = 1$. Then

$$\begin{aligned} p^* \cdot z(p) &= (p^* - p) \cdot (z(p) - z(p^*)) \\ &= (p_1^* - p_1)(z_1(p) - z_1(p^*)) > 0. \end{aligned}$$

The first equality uses Walras's Law (giving us that $p \cdot z(p) = 0$) and the fact that p^* is a Walrasian equilibrium (so that $z(p^*) = 0$). The second equality uses the normalization $p_2 = p_2^* = 1$. The inequality follows from the gross substitutes property: $p_1 > p_1^*$ implies $z_1(p) < z_1(p^*)$ and $p_1 < p_1^*$ implies $z_1(p) > z_1(p^*)$.■

With this Lemma, we can prove that the tatonnement process converges to the unique (up to normalization) Walrasian equilibrium price vector p^*.

Proposition 2. Suppose that the aggregate excess demand function $z(\cdot)$ is satisfies the gross substitutes property and that p^* is a Walrasian equilibrium price vector. Then the price-adjustment process $p(t)$ defined by $dp(t)/dt = \alpha z(p(t))$, with $\alpha > 0$, converges to p^* for any initial condition $p(0)$.

Proof of Proposition 2. To prove this result, we will show that the squared distance between $p(t)$ and p^* decreases monotonically in t. Let $D(p) = \frac{1}{2}\sum_{l \in \mathcal{L}} (p_l - p_l^*)^2$ denote the distance between p and p^*. Then

$$\begin{aligned}
\frac{dD(p(t))}{dt} &= \sum_{l \in \mathcal{L}} (p_l(t) - p_l^*) \frac{dp_l(t)}{dt} \\
&= \alpha \sum_{l \in \mathcal{L}} (p_l(t) - p_l^*) z_l(p(t)) \\
&= -\alpha p^* \cdot z(p) \leq 0,
\end{aligned}$$

where the third equality uses Walras's law. By the previous lemma, the last inequality is strict unless p is collinear with p^*. Since $D(p(t))$ is monotonic and bounded, it must converge to some value $\delta \geq 0$. If $\delta = 0$, we are done. If

$\delta > 0$, then there is a contradiction, because continuity of aggregate demand implies that $p^* \cdot z(p(t))$ is bounded away from 0 for all $p(t)$ bounded away from p^*.■

Finally, economies with the gross substitutes property have nice comparative statics. Any change that raises excess demand for commodity k will increase its equilibrium price. As an example, suppose there are two commodities and normalize $p_2 = 1$. Suppose also that commodity 1 is a normal good for all consumers. Consider an increase in the aggregate endowment for commodity 2. For any price p_1, this will increase aggregate demand for commodity 1 and hence increase $z_1(\cdot, 1)$.

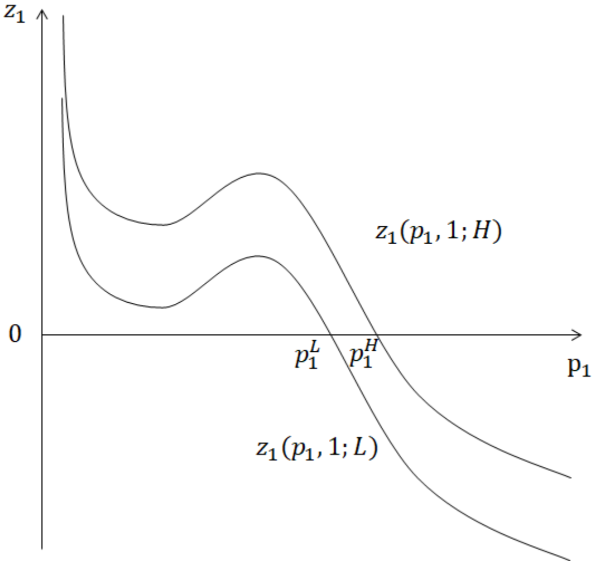


Figure 14: Comparative statics

Figure 14 compares the aggregate excess demand functions for two economies: one $(z_1(\cdot, 1; L))$ with a low aggregate endowment of commodity 2 and one $(z_1(\cdot, 1; H))$ with a high aggregate endowment of commodity 2. The curve $z_1(\cdot, 1; H)$ lies above $z_1(\cdot, 1; L)$ and because it is continuous and crosses zero once, the new equilibrium price vector must have a higher price for commodity 1.

1.6 Empirical Content of GE

As we just saw, whether there is a unique Walrasian equilibrium and whether Walrasian equilibria are stable depended critically on the structure of the economy's aggregate excess demand function $z(\cdot)$.

What do we know in general about the structure of aggregate excess demand? We proved that under assumptions $(A1)$, $(A2')$, $(A3)$, and $(A4)$ about consumer preferences and endowments that $z(\cdot)$ is continuous, homogeneous of degree zero in p, it satisfies Walras's Law, and $\lim_{p \to 0} z(p) \to \infty$. But, as Sonnenschein (1973) showed for the case of two commodities, and Mantel (1974) and Debreu (1974) showed more generally, the assumption of consumer maximization alone imposes no further restrictions on $z(\cdot)$. This is a very negative result, since it implies that even if we observe an economy in a Walrasian equilibrium with price vector p, it is possible for the same economy to have an arbitrary number of Walrasian equilibria with arbitrary stability properties.

Theorem 5 (Sonnenschein-Mantel-Debreu Theorem). For any closed and bounded set of positive prices $\mathcal{P} \subseteq \mathbb{R}^L_{++}$ and any function $f : \mathcal{P} \to \mathbb{R}^L$ satisfying continuity, homogeneity of degree 0, and Walras's Law, there exists an exchange economy with L consumers with continuous, strictly convex, and monotone preferences whose aggregate excess demand function coincides with f on $\mathcal{P}$.

We omit the proof here. See MWG Chapter 17.E for a proof in the $L = 2$ case and a discussion about the more general proof. Roughly speaking, the structure of the proof begins with a candidate excess demand function $f(p)$ that is continuous, homogeneous of degree 0, and satisfies Walras's Law and reverse engineers a set of consumer preferences and endowments that generate $f(p)$ as the aggregate excess demand function. The ability to do so requires a lot of flexibility in specifying consumer preferences that feature potentially strong income effects as well as the ability to specify consumers' endowments. A common interpretation of this theorem is that "anything goes" in general equilibrium theory. That is, without making strong assumptions on preferences: (i) pretty much any comparative statics result could be obtained in a general equilibrium model, and (ii) general equilibrium theory has essentially no empirical content. This is not quite right, though, as we will now see.

Brown and Matzkin (1996) prove an important result showing that if an economist is able to observe endowments as well as prices, then the Walrasian model is in principle testable. That is, there are endowment and price pairs

$\left(p, (\omega_i)_{i\in\mathcal{I}}\right)$ and $\left(p', (\omega_i')_{i\in\mathcal{I}}\right)$ such that if p is a Walrasian equilibrium price vector given a fixed set of consumers with endowments $(\omega_i)_{i\in\mathcal{I}}$, then if the same set of consumers instead had endowments $(\omega_i')_{i\in\mathcal{I}}$, p' could not be a Walrasian equilibrium price vector.

Theorem 6 (Brown-Matzkin Theorem). There exists price-endowment pairs $\left(p, (\omega_i)_{i\in\mathcal{I}}\right)$ and $\left(p', (\omega_i')_{i\in\mathcal{I}}\right)$ such that there do not exist monotone preferences $(u_i)_{i\in\mathcal{I}}$ such that p is a Walrasian equilibrium price vector for the exchange economy $(u_i, \omega_i)_{i\in\mathcal{I}}$ and p' is a Walrasian equilibrium price vector for the exchange economy $(u_i, \omega_i')_{i\in\mathcal{I}}$.

Proof of Theorem 6. We can prove this theorem in the case of two consumers and two commodities. Consider the two Edgeworth boxes in Figure 15. Because p is a Walrasian equilibrium price vector given endowment ω, consumer 1 must weakly prefer some bundle on the segment A to any bundle on the segment B. By monotonicity, for every point on the segment A', there is some point on B that consumer 1 strictly prefers. There is therefore some bundle on A that is preferred by consumer 1 to every bundle on A'. If p' is a Walrasian equilibrium price vector given ω', we have a contradiction: every bundle on A is available to consumer 1 at prices p', yet she chooses a bundle on A'.■

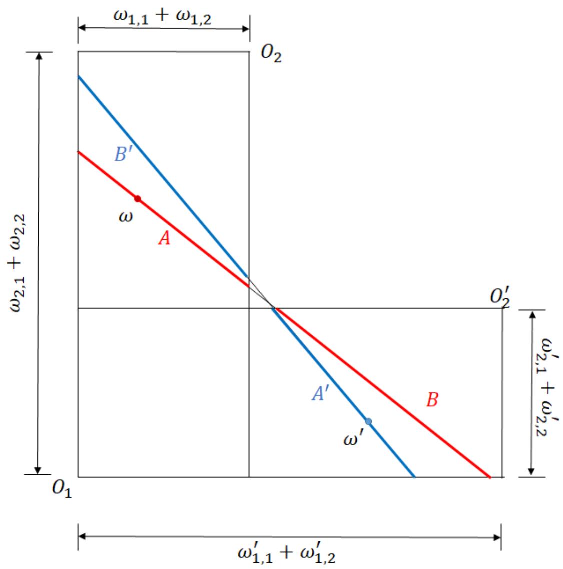

$\omega_{1,1}+\omega_{1,2}$
O_2
$\omega_{2,1}+\omega_{2,2}$
B'
ω
A
O_2'
$\omega'_{2,1}+\omega'_{2,2}$
A'
B
ω'
O_1
$\omega'_{1,1}+\omega'_{1,2}$

Figure 15: Brown-Matzkin theorem

The Brown-Matzkin theorem shows that, in order to construct the arbitrary excess demand functions that the proof of the Sonnenschein-Mantel-Debreu theorem requires, you really need the flexibility in specifying arbitrary endowments in addition to flexibility in specifying preferences. It also illustrates a more general point that, even if at its highest level, a theory imposes little structure on endogenous variables, imposing more structure on the theory typically imposes more structure on its implications.

Chapter 2

Foundations of General Equilibrium Theory

When we talked about the normative properties of the GE model in the first week of class—in particular, the first welfare theorem—we were focused on questions about *what* is true in equilibrium. We did not really address the question of *why* it is that Walrasian equilibrium allocations are Pareto optimal, because in some sense, the framework itself is ill-equipped to answer this question. In order to get a better sense for why Walrasian equilibrium allocations are Pareto efficient, and in order to get a better sense for when the GE framework is an appopriate way of viewing the world, we will have to take a step outside the model in order to provide some foundations for the model itself.

We will ask two sets of questions. First, when might we expect Walrasian equilibrium allocations to arise? Second, when and why might we expect Pareto-optimal allocations to arise? To address these questions, we will consider two alternative foundations for Walrasian equilibrium, and in both, the answer to this question will be: when individuals in the economy are "small."

What is potentially problematic about Walrasian equilibrium as a description of the economy is that prices are endogenous variables, but they are not the explicit choices of anyone in the economy. In reality, individuals set prices—they bid in auctions, they post prices in their stores, they negotiate prices with their suppliers. Setting prices are individual decisions. One of the main premises of GE is that, when individuals are small relative to the economy, "market forces" pin down the prices at which trade occurs, and although it may be possible, it would be unwise for individuals to choose any other prices or any other consumption bundles. The question is: what are "market forces?"

Providing microfoundations for GE theory boils down to providing an answer to the question, "Under what conditions do small individuals lack market power, in the sense that they are forced to trade only at competitive prices?" There are two main approaches we will consider here. The first is the *cooperative game theory approach* in which the primitives of the model remain the agents' endowments and preferences, and the process of price-setting and trade is still specified only implicitly. Under this approach, however, the solution concept we will be using does not directly involve prices. Yet as the economy becomes large, consumers will receive the same allocations they would receive in a Walrasian equilibrium.

The second approach we will consider is the *non-cooperative game theory approach* in which we will explicitly model price-setting and trade and think about the (Nash) equilibria of the resulting trading processes. Consumers will have actions that can directly affect the prices they and other consumers pay for different commodities, and therefore equilibria will generically be inefficient. In the limit as the economy becomes large, however, consumers' actions will have little effect on prices, and equilibrium consumption choices will converge to Walrasian equilibrium allocations.

A benefit of the non-cooperative approach relative to the cooperative approach is that Pareto optimality will arise as a result rather than as a maintained assumption. We will therefore be able to develop some deeper intuition for why, exactly, Walrasian equilibrium allocations are Pareto optimal. We will then conclude this section with a brief discussion of who gets what in equilibrium and how under the notion of *competitive equilibrium* in which consumers' impact on prices is miniscule, consumers receive exactly what they contribute to the economy. A common theme in these approaches is that while Walrasian equilibrium is not necessarily a good description of small-numbers interactions, it may be a reasonable description of large-numbers interactions.

2.1 The Cooperative Approach

Going back to his 1881 classic, *Mathematical Psychics*, Edgeworth proposed that in economies with a small number of individuals, the outcome might be "indeterminate." We saw an example of this in the first week when we looked at Edgeworth boxes—with two consumers, Edgeworth believed that the only prediction we could reasonably make is that the final allocation would lie on the contract curve: the set of Pareto optimal allocations that are preferred by each consumer to her endowment. But he also conjectured that as the number of consumers grows, the scope for contracting among different consumers grows, and the resulting contract curve shrinks until it reaches only the set of Walrasian equilibrium allocations.

In the 20th century, economists formalized a version of this argument in what is known as the *core convergence theorem.* In order to describe what the core convergence theorem is, we will first have to define what the *core* is. The idea of the core is that it is the set of allocations for which no group of consumers can get together and trade with each other and do strictly better. Formally, consider a pure exchange economy $\mathcal{E}$ with I consumers whose preferences are continuous, strictly convex, and strongly monotone. We will define the core by defining what it is not. We will say that a coalition $\mathcal{S} \subseteq \mathcal{I}$ of consumers *blocks* an allocation if its members can all do strictly better by trading among themselves. In the case of $\mathcal{I} = \{1, 2\}$ that we considered in the Edgeworth box, any allocation that is not in the Pareto set is blocked by the coalition $\{1, 2\}$, and any allocation in the Pareto set but not on the contract curve is blocked either by coalition $\{1\}$ or by coalition $\{2\}$.

Definition 5. A coalition $\mathcal{S} \subseteq \mathcal{I}$ **blocks** the allocation $x^* = (x_1^*, \ldots, x_I^*) \in \mathbb{R}_+^{LI}$ if there exists another allocation such that:

1. $u_i(x_i) > u_i(x_i^*)$ for all $i \in \mathcal{S}$, and
2. $\sum_{i \in \mathcal{S}} x_i \leq \sum_{i \in \mathcal{S}} \omega_i$.

The core is the set of feasible unblocked allocations.

Definition 6. A feasible allocation x^* is in the **core** if it is not blocked by any coalition. The core is therefore the set of unblocked feasible allocations.

In terms of the Edgeworth box example, the core corresponds to the contract curve, since all other allocations are blocked by some coalition. The core convergence theorem provides conditions under which, when the economy grows large, the set of *core* allocations coincides with the set of *Walrasian equilibrium* allocations. We will break this claim up into two parts. First, we will show that any Walrasian equilibrium allocation is in the core. Then, we will show that any allocation that remains in the core as the economy grows large is a Walrasian equilibrium allocation.

Proposition 3. Any Walrasian equilibrium allocation is in the core.

Proof of Proposition 3. Let $\left(p^*, (x_i^*)_{i \in \mathcal{I}}\right)$ be a Walrasian equilibrium. Suppose $(x_i^*)_{i \in \mathcal{I}}$ is not in the core. Then there is some coalition $\mathcal{S} \subseteq \mathcal{I}$ that can block x^* with some other feasible allocation $\hat{x}$. Then $p^* \cdot \hat{x}_i > p^* \cdot \omega_i$ for all $i \in \mathcal{S}$ by consumer optimality. Since this holds for all $i \in \mathcal{S}$, we must also have $p^* \cdot \left(\sum_{i \in \mathcal{S}} \hat{x}_i\right) > p^* \cdot \left(\sum_{i \in \mathcal{S}} \omega_i\right)$. Since $p^* \geq 0$, this implies that $\sum_{i \in \mathcal{S}} \hat{x}_{l,i} > \sum_{i \in \mathcal{S}} \omega_{l,i}$ for some commodity l. But this means that $\hat{x}$ was not feasible to begin with.■

In order to establish the other direction of the core convergence theorem, we will have to define formally what we mean when we say that an economy grows large. As we know from the Edgeworth box example, when there are only two consumers, not every core allocation is a Walrasian equilibrium allocation. Whether this result remains true as we add more consumers to the economy depends on *how* we add more consumers to the economy. For example, if consumers 1 and 2 only care about their consumption of pens and pencils, and they are endowed with pens and pencils and nothing else, then if we add a bunch of other consumers who care only about their consumption of paper clips and have an endowment of paper clips and nothing else, then this will not do anything to make the terms of trade between consumers 1 and 2 more competitive.

If instead, we "grow the economy" by adding more consumers like consumer 1 (i.e., consumers who have the same preferences and endowment as consumer 1) and adding more consumers like consumer 2, then core allocations do begin to look more like Walrasian equilibrium allocations. Roughly speaking, the reason why is that if any particular consumer is getting a "good deal" from the rest of the consumers at a particular allocation, then the other consumers would prefer to cut her out of the deal and redistribute her net trade among themselves. This may not work when there are only a couple consumers in the economy because the excluded consumer may be hard to replace.

In order to formalize this argument, suppose there are H **types** of consumers $h \in \mathcal{H} = \{1, \ldots, H\}$. A type-$h$ consumer has preferences u_h and endowment ω_h. For each integer $N > 0$, we consider the **N-replica economy**, which is a pure exchange economy consisting of $I_N \equiv N \cdot H$ consumers, N of each type. We will refer to an allocation in which consumers of the same type consume the same consumption bundle as an **equal-treatment allocation**. The next lemma establishes that any core allocation of the N-replica economy is an equal-treatment allocation. Denote by $x_{h,n}$ the allocation of the n^{th} consumer of type h.

Lemma 5. Suppose x is in the core of the N-replica economy, for some $N > 0$. Then all consumers of the same type receive the same allocation: $x_{h,n} = x_{h,m}$ for all $n, m \leq N$ and all types $h \in \mathcal{H}$.

Proof of Lemma 5. We will proceed by way of contradiction. Suppose x is in the core of the N-replica economy for some $N > 0$, but for some type of consumer—without loss of generality, say type 1—not all consumers of that type receive the same allocation. We will want to show that in fact, such an allocation is not in the core. In particular, we will show that the coalition consisting of the worst-off consumer of every type can block the allocation x.

To see why this is true, let $\hat{x}_h = \frac{1}{N} \sum_n x_{h,n}$ denote the average allocation of type-h consumers. Without loss of generality, suppose that it is consumer number 1 of each type h who is worst off within type h. By strict convexity of preferences, $u_h(\hat{x}_h) \geq u_h(x_{h,1})$ for all h, and $u_1(\hat{x}_1) > u_1(x_{1,1})$. The coalition $\{(1,1),\ldots,(H,1)\}$ can attain consumption vector $(\hat{x}_1,\ldots,\hat{x}_H)$ for its members, since feasibility implies

$$\sum_{h\in\mathcal{H}}\hat{x}_h = \frac{1}{N}\sum_{h\in\mathcal{H}}\sum_{n=1}^{N}x_{h,n} \leq \frac{1}{N}\sum_{h\in\mathcal{H}}\sum_{n=1}^{N}\omega_h = \sum_{h\in\mathcal{H}}\omega_h.$$

Finally, continuity and strong monotonicity of preferences imply that the consumption vector $(\hat{x}_1,\ldots,\hat{x}_H)$ can be perturbed to satisfy $u_h(\hat{x}_h) > u_h(x_{h,1})$ for all h, so the strict inequalities required to apply the definition of blocking are satisfied.■

This Lemma shows that any core allocation for an N-replica economy takes the form of a **type allocation** $(x_1,\ldots,x_H) \in \mathbb{R}_+^{LH}$, where each consumer of type h receives allocation x_h. Let $\mathcal{C}_N \subseteq \mathbb{R}_+^{LH}$ be the set of core allocations in the N-replica economy. Note that the set of core allocations shrinks as we replicate the economy: $\mathcal{C}_{N+1} \subseteq \mathcal{C}_N$ for all N. This is because any type allocation that is blocked by some coalition in the N-replica economy will be blocked by exactly the same coalition in the $N+1$-replica economy. At the same time, from Proposition 3, we know that the the set of Walrasian equilibrium allocations is independent of N and is always contained in $\mathcal{C}_N$. Debreu and Scarf (1963) proved that as $N \to \infty$, the set $\mathcal{C}_N$ shrinks to exactly the set of Walrasian equilibrium allocations. The version of the theorem we will prove will rely on two additional assumptions about preferences and endowments, although these assumptions can be relaxed.

Assumption A1' (continuous differentiability). For all consumers of type $h \in \mathcal{H}$, u_h is continuously differentiable.

Assumption A4' (interiority). For each $h \in \mathcal{H}$, ω_h is strictly preferred to any consumption bundle x_h that is not strictly positive.

Theorem 7 (Core Convergence Theorem). Suppose $\mathcal{E}$ satisfies $(A1')\,,(A2')\,,(A3)\,,(A4')$. If $x \in \mathcal{C}_N$ for all N, then x is a Walrasian equilibrium allocation.

Proof of Theorem 7. At a high level, the proof of this theorem first argues that if $x \in \mathcal{C}_N$ for all N, then it is Pareto-optimal, which means that marginal rates of substitutions are equal across consumers and proportional to a price vector that will be used to construct a Walrasian equilibrium. It then argues that if at this price vector, some type of consumer is getting a "good deal" in that they consuming a bundle that is more expensive than their endowment, then $N-1$ consumers of this type along with all the other consumers in the economy can form a blocking coalition. This means that no type of consumer can be getting a good deal if $x \in \mathcal{C}_N$ for all N. The proof concludes with an argument that if no consumers are getting a good deal at $x \in \mathcal{C}_N$ for all N, then x can be decentralized as a Walrasian equilibrium allocation.

Step 1. Pareto-optimal allocations equate marginal rates of substitution across consumers and can be used to construct candidate prices.

Take an $x^* \in \mathcal{C}_N$ for all N. Since x^* is in the core, it is a Pareto-optimal allocation. Assumptions $(A1')$ and $(A4')$ ensure that at x^*,

$$\frac{\partial u_h / \partial x_{l,h}}{\partial u_h / \partial x_{l',h}} = \frac{\partial u_{h'} / \partial x_{l,h'}}{\partial u_{h'} / \partial x_{l',h'}} \text{ for all } h, h', l, l'.$$

Construct a price vector p^* for which $p_1^* = 1$, and

$$p_l^* = \frac{\partial u_h / \partial x_{l,h}}{\partial u_h / \partial x_{1,h}} \text{ for any } h,$$

so that relative prices match relative marginal utilities. We will now argue that (p^*, x^*) is a Walrasian equilibrium.

Step 2. No consumer types are getting a "good deal" at p^*.

Suppose that type-1 consumers are getting a "good deal" in the sense that their consumption is worth more than their endowment at prices p^*: $p^* \cdot x_1^* > p^* \cdot \omega_1$. We want to show that if this is the case, then x^* is not, in fact, in $\mathcal{C}_N$ for all N. To see why this is the case, note the marginal utility to any consumer type h of consuming an additional ε amount of consumer 1's net trade, $x_1^* - \omega_1$, is, to first order,

$$\varepsilon \sum_{l \in \mathcal{L}} \frac{\partial u_h}{\partial x_{l,h}} (x_{l,1} - \omega_{l,1}).$$

Since $p^* \cdot (x_1^* - \omega_1) > 0$, and the vector $(\partial u_h / \partial x_{1,h}, \ldots, \partial u_h / \partial x_{L,h})$ is proportional to p^*, this marginal utility is strictly positive. For $\varepsilon > 0$ sufficiently small, therefore, each consumer type h strictly prefers consuming $x_h^* + \varepsilon (x_1^* - \omega_1)$ to consuming x_h^*.

Now, consider allocation x^* in the N-replica economy. Suppose the coalition $\mathcal{S}$ consisting of everyone except a single type-1 consumer proposes an allocation that gives each coalition member of type h consumption $\hat{x}_h = x_h^* + \frac{1}{NH-1}(x_1^* - \omega_1)$. This allocation $\hat{x}$ is feasible for the coalition (you can check MWG, p. 658 for the argument for feasibility). Moreover, by the argument in the previous paragraph, if N is sufficiently large, $\hat{x}$ is strictly preferred to x^* by every coalition member. The coalition $\mathcal{S}$ therefore blocks the allocation x^*, so x^* is not in $\mathcal{C}_N$ for N sufficiently large. This contradicts the hypothesis that $p^* \cdot x_1^* > p^* \cdot \omega_1$, so it must be the case that no consumer types are getting a "good deal."

Step 3: Show that (p^*, x^*) is a Walrasian equilibrium.

From the previous step, we know that x_h^* is affordable for type h at prices p^* for all types h: $p^* \cdot x_h^* \leq p^* \cdot \omega_h$ for all $h \in \mathcal{H}$. The bundle x_h^* also satisfies consumer optimality. This is because under our interiority, differentiability, and convexity assumptions, each consumer type h will choose a consumption bundle that equates $\frac{\partial u_h}{\partial x_{l,h}}/p_l^*$ across commodities and therefore will optimally choose x_h^* at prices p^*.

Finally, since x^* is Pareto-optimal, it must also be feasible: $N\sum_{h\in\mathcal{H}} x_h^* \leq N\sum_{h\in\mathcal{H}} \omega_h$. Since preferences are monotone, this inequality must hold with equality, so the market-clearing condition is also satisfied. The vector (p^*, x^*) is therefore a Walrasian equilibrium.■

The core convergence theorem is an important result that is probably the best-known statement of the idea that large markets are approximately competitive. Note that there are no prices in the notion of the core. Yet what the core convergence theorem is saying is that in a sufficiently large economy, any allocation in the core corresponds to exactly what consumers would consume at equilibrium prices in a Walrasian equilibrium. The theorem itself has a number of shortcomings, however.

First, the notion of a replica economy is extreme. We typically think of each individual as being unique, yet the thought experiment the core convergence theorem carries out requires that there are, in the limit, infinitely many people who have exactly the same preferences and endowments as you. Second, the theorem itself is not an approximation result—it does not say that for any finite N, any allocation in the core is *approximately* a Walrasian equilibrium allocation, since it does not say anything about distance. There is a large literature at the intersection of cooperative game theory and general equilibrium theory that tries to extend this result into something that is more convincing. One branch (following Arrow and Hahn, 1971, and others) relaxes the assumption of exact replication and tries to say something about core allocations in large but finite economies. Another branch (following Aumann, 1964) instead looks directly at economies with a continuum of consumers, for which the core convergence theorem provides an exact equivalence between core allocations and Walrasian equilibrium allocations.

2.2 The Non-Cooperative Approach

The cooperative approach imposes no structure on the underlying trading institutions and as a result, it has little to say about how prices are determined and under what conditions they are likely to correspond to Walrasian equilibrium prices. In contrast, under the non-cooperative approach, individual consumers make decisions that "aggregate up" to determine prices.

Suppose there are I consumers, a set $\mathcal{P} \subseteq \mathbb{R}^L$ of possible price vectors, and a set $\mathcal{A}$ of **market actions**. Each consumer $i \in \mathcal{I}$ has a set $\mathcal{A}_i \subset \mathcal{A}$ and an endowment vector $\omega_i \in \mathbb{R}^L$. For each $a_i \in \mathcal{A}_i$ and $p \in \mathcal{P}$, a **trading rule** assigns a net trade vector $g(a_i; p) \in \mathbb{R}^L$ to consumer i, satisfying $p \cdot g(a_i; p) = 0$. Given a vector of market actions $a = (a_1, \ldots, a_I)$, a market-clearing process generates a price vector $p(a) \in \mathcal{P}$. Throughout, we will assume each i has a utility function of the form $u_i(g(a_i; p) + \omega_i)$. An equilibrium of the resulting game is just a Nash equilibrium.

Definition 7. The profile $a^* = (a_1^*, \ldots, a_I^*)$ of market actions is a **trading equilibrium** if, for every consumer $i \in \mathcal{I}$,

$$u_i(g(a_i^*; p(a^*)) + \omega_i) \geq u_i\left(g\left(a_i; p\left(a_i, a_{-i}^*\right)\right) + \omega_i\right) \text{ for all } a_i \in \mathcal{A}_i.$$

We will consider a particular trading rule referred to as Shapley and Shubik's (1977) **trading posts**. It is not particularly realistic, but it does form a complete general equilibrium model in which all consumers interact strategically. Suppose there are I consumers and L commodities. Commodity L, which we will call "money," is treated differently from the other commodities, and we normalize its price to 1. For each of the other $L-1$ commodities, there is a *trading post* at which consumers can exchange money for the commodity.

At each trading post $l \leq L-1$, each consumer i places bids $a_{l,i} = \left(a'_{l,i}, a''_{l,i}\right) \in \mathbb{R}^2_+$. The first value, $a'_{l,i}$, is interpreted as the amount of commodity l that consumer i is willing to put up for sale in exchange for money. The second value, $a''_{l,i}$ is the amount of money that she puts up in exchange for commodity l. These bids must therefore satisfy $a'_{l,i} \leq \omega_{l,i}$, and $\sum_{l \leq L-1} a''_{l,i} \leq \omega_{L,i}$. Given the consumers' bids, the price of commodity l is set to be equal to the total amount spent on commodity l divided by the total quantity of commodity l supplied:

$$p_l = \frac{\sum_{i \in \mathcal{I}} a''_{l,i}}{\sum_{i \in \mathcal{I}} a'_{l,i}}.$$

Each consumer i receives allocation $x_{l,i} = g_l\left(a_i; p\right) + \omega_{l,i}$, where

$$g_l\left(a_i; p\right) = \frac{a''_{l,i}}{p_l} - a'_{l,i}$$

for all $l \leq L-1$ and $x_{L,i} = \omega_L - \sum_{l=1}^{L-1} a''_{l,i}$.

If there is a large number of consumers trading each commodity, then each consumer's bids would have a negligible effect on prices, and each consumer's allocation will be arbitrarily close to the solution to their problem

$$\max_{x_i \in \mathbb{R}^L_+} u_i(x_i) \text{ s.t. } p \cdot x_i \leq p \cdot \omega_i.$$

Thus, even though prices *are* determined as the aggregation of individual consumers' actions, when the economy is sufficiently large, each *individual* consumer's actions have no effect on prices. Under this approach, price-taking behavior is therefore a result rather than an assumption. We will refer to the resulting equilibrium as a *competitive equilibrium.*

One important difference between the cooperative approach and the non-cooperative approach to the foundations of GE theory is that under the cooperative approach, the allocations we considered were always Pareto optimal. In contrast, under the non-cooperative approach, allocations are *not* Pareto efficient for any finite market size. When the size of the economy grows does the set of equilibrium allocations become approximately Pareto optimal. Only the non-cooperative approach can, therefore, really tell us anything about *why* Walrasian equilibrium allocations are Pareto optimal.

2.3 Who Gets What? The No-Surplus Condition

This section concludes our discussion of the competitive foundations of general equilibrium theory. In particular, we will ask whether Walrasian equilibria can be characterized by the idea that consumers get exactly what they contribute to the welfare of society. To answer this question, we will consider a special class of preferences in which the notion of the *welfare of society* is well-defined. In particular, suppose there are H types of consumers, $h \in \mathcal{H} = \{1, \dots, H\}$, and each type of consumer is endowed with ω_h and has *quasi-linear preferences.*

Assumption QL (quasilinearity). For each type $h \in \mathcal{H}$, there is a concave, differentiable, strictly increasing function $v_h \left(x_{1,h}, \dots, x_{L-1,h} \right)$ such that type h preferences are $u_h \left(x_h \right) = v_h \left(x_{1,h}, \dots, x_{L-1,h} \right) + x_{L,h}$, where $x_h \in \mathbb{R}_+^{L-1} \times \mathbb{R}$.

When consumers have quasilinear preferences, commodity L is what is referred to as the **money commodity**. It is a commodity for which all consumers have the same marginal utility and which consumers can consume any (positive or negative) amount of. The assumption of quasilinear preferences allows for cardinal measures of individuals' private rewards and their contribution to social welfare.

An economy is defined by a profile $(I_1, \dots, I_H)$ of consumers of the different types, for a total of $I = \sum_{h \in \mathcal{H}} I_h$ consumers. For any economy, we can define the **social welfare**, $V \left(I_1, \dots, I_H \right)$, as the solution to the following problem:

$$V \left(I_1, \dots, I_H \right) = \max_{\left(x_h \right)_{h \in \mathcal{H}}} \sum_{h \in \mathcal{H}} I_h u_h \left(x_h \right)$$

subject to feasibility: $\sum_{h \in \mathcal{H}} I_h x_h \leq \sum_{h \in \mathcal{H}} I_h \omega_h$ and $x_{l,h} \geq 0$ for all $l \in$

$\{1, \ldots, L-1\}$ and for all h. This function is homogeneous of degree one in its arguments, so we can describe the economy in terms of its per-capita social welfare $V(I_1/I, \ldots, I_H/I) = V(I_1, \ldots, I_H)/I$, and therefore if we extend the model to one in which there are a continuum of consumers, with mass $\mu_h \geq 0$ of type $h \in \mathcal{H}$ with $\sum_{h \in \mathcal{H}} \mu_h = 1$, we can write $\mu = (\mu_1, \ldots, \mu_H)$ and

$$V(\mu) = \max_{(x_h)_{h \in \mathcal{H}}} \sum_{h \in \mathcal{H}} \mu_h u_h(x_h) \tag{1}$$

subject to feasibility: $\sum_{h \in \mathcal{H}} \mu_h x_h \leq \sum_{h \in \mathcal{H}} \mu_h \omega_h$ and $x_{l,h} \geq 0$ for all $l \in \{1, \ldots, L-1\}$.

Given a continuum population of consumers, we can define a consumer of type h's **marginal contribution to social welfare** as $\partial V(\mu)/\partial \mu_h$. We will say that a feasible allocation $(x_h^*)_{h \in \mathcal{H}}$ is a **no-surplus allocation** if

$$u_h(x_h^*) = \frac{\partial V(\mu)}{\partial \mu_h} \text{ for all } h \in \mathcal{H}.$$

In other words, at a no-surplus allocation, each consumer is receiving in utility exactly what she contributes to social welfare. With this definition in mind, we can state the no-surplus characterization of Walrasian equilibrium (Ostroy, 1980, Makowski and Ostroy, 1995).

Theorem 8 (No-Surplus Characterization). For any continuum population $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_H) >> 0$, a feasible allocation $(x_1^*, \ldots, x_H^*) >> 0$ is a no-surplus allocation if and only if it is a Walrasian equilibrium allocation.

Proof of Theorem 8. The structure of the proof is as follows. We will show that if $(x_h^*)_{h\in\mathcal{H}}$ is a no-surplus allocation, then it solves (1). We will then show that if $(x_h^*)_{h\in\mathcal{H}}$ solves (1), then $(x_h^*)_{h\in\mathcal{H}}$ is a Walrasian equilibrium allocation for a suitable price vector p^*. Finally, we will show that if $(x_h^*)_{h\in\mathcal{H}}$ is a Walrasian equilibrium allocation, then it is a no-surplus allocation.

Step 1: $(x_h^*)_{h\in\mathcal{H}}$ is no-surplus $\Rightarrow$ $(x_h^*)_{h\in\mathcal{H}}$ solves (1).

Suppose $(x_h^*)_{h\in\mathcal{H}}$ is a no-surplus allocation. We know that the function $V(\bar{\mu})$ is homogeneous of degree one in $\bar{\mu}$, so by Euler's formula, we can write

$$V(\bar{\mu}) = \sum_{h\in\mathcal{H}} \bar{\mu}_h \frac{\partial V(\bar{\mu})}{\partial \mu_h} = \sum_{h\in\mathcal{H}} \bar{\mu}_h u_h(x_h^*),$$

where the last equality used the fact that $(x_h^*)_{h\in\mathcal{H}}$ is a no-surplus allocation. This implies that $(x_h^*)_{h\in\mathcal{H}}$ is a solution to the social welfare-maximization problem for $\mu = \bar{\mu}$.

Step 2: $(x_h^*)_{h\in\mathcal{H}}$ solves (1) $\Rightarrow$ $(x_h^*)_{h\in\mathcal{H}}$ is a WE allocation.

Suppose now that $(x_h^*)_{h\in\mathcal{H}}$ is a feasible allocation that yields social welfare $V(\bar{\mu})$. Denote by p_l^*, $l = 1, \ldots, L$, the values of the Lagrange multipliers for commodity-l feasibility constraint, $\sum_{h\in\mathcal{H}} \bar{\mu}_h (x_{l,h} - \omega_{l,h}) \leq 0$, in the social-welfare-maximization problem. Because $u_h(\cdot)$ is quasilinear for all $h \in \mathcal{H}$, we will have $p_L^* = 1$ and $p_l^* = \partial u_h(x_h^*) / \partial x_{l,h}$ for all $l \in \{1, \ldots, L-1\}$ and for all $h \in \mathcal{H}$. It follows then that if we let $p^* = (p_1^*, \ldots, p_L^*)$, then $\left(p^*, (x_h^*)_{h\in\mathcal{H}}\right)$ is a Walrasian equilibrium.

Step 3: $(x_h^*)_{h\in\mathcal{H}}$ is a WE allocation $\Rightarrow$ $(x_h^*)_{h\in\mathcal{H}}$ is no-surplus.

Finally, we can apply the envelope theorem to (1) to get

$$\frac{\partial V\left(\bar{\mu}\right)}{\partial \mu_h} = u_h\left(x_h^*\right) + p^* \cdot \left(\omega_h - x_h^*\right).$$

Since $\left(x_h^*\right)_{h\in\mathcal{H}}$ is consumer-optimal given prices p^*, by Walras's law, the second term is zero. We therefore have that $\partial V\left(\bar{\mu}\right)/\partial \mu_h = u_h\left(x_h^*\right)$, so $\left(x_h^*\right)_{h\in\mathcal{H}}$ is a no-surplus allocation.■

Viewed in light of the no-surplus characterization of Walrasian equilibrium, we can finally develop some intuition for the first welfare theorem result that Walrasian equilibrium allocations are Pareto optimal. If, at the margin, each consumer is receiving exactly what she contributes to society's welfare, then in some sense, the rest of society is indifferent to her presence. Since each consumer is not affecting the welfare of the rest of society, of course each consumer doing the best she can—which she is, by the consumer optimality condition of Walrasian equilibrium—is going to lead to a result that is best for society.

It is important to realize that when there are a finite number of individuals in society, there generically do not exist any no-surplus allocations. The reason for this is that it is typically impossible to give each consumer the full extent of her marginal contribution while maintaining feasibility. For example, when there are only two consumers in the economy, each consumer's contribution to social welfare is equal to the utility she would get if she consumes her endowment plus the *entire gains from trade*, and we cannot simultaneously give both consumers the entire gains from trade. This means that in smaller economies, Walrasian equilibrium allocations generically are not no-surplus allocations.

Chapter 3

Extensions of the GE Framework

3.1 Firms and Production in General Equilibrium

So far in this class, we have focused on pure exchange economies. In doing so, we have assumed that all the commodities in the economy come essentially from nowhere. In other words, we have completely abstracted away from the supply side of the economy. The GE framework can be readily extended to allow for firms and productions as long as two conditions are satisfied: (1) firms' production technologies do not exhibit increasing returns to scale, and (2) firms are price-takers. In this section, we will describe how to extend the GE framework to allow for production and we will show that versions of the welfare theorems and the existence theorem hold. We will then consider some simple examples and conclude with a result that shows that in this framework, we can think of the entire supply side of the economy as a single firm.

The Model There are I consumers $i \in \mathcal{I}$ with utility functions $(u_i)_{i \in \mathcal{I}}$ defined over the consumption of L commodities $l \in \mathcal{L}$, and there are J firms $j \in \mathcal{J}$. Each firm possesses a production set $\mathcal{Y}_j \in \mathbb{R}^L$. The production set $\mathcal{Y}_j$ describes a set of feasible production plans: if $y_j = (y_{1,j}, \ldots, y_{L,j}) \in \mathcal{Y}_j$, then $y_{l,k} < 0$ means that commodity l is being used as an input, and $y_{l,k} > 0$ means that commodity l is being produced as an output. The firms are owned by the households. **Consumer** i's **ownership share of firm** j is a $\theta_{i,j} \in [0,1]$. A **production economy** is then a collection $\mathcal{E} = \left(\left(u_i, \omega_i, (\theta_{i,j})_{j \in \mathcal{J}}\right)_{i \in \mathcal{I}}, (\mathcal{Y}_j)_{j \in \mathcal{J}}\right)$ of consumer preferences, consumer endowments, ownership shares, and production sets. Firm j takes prices $p \in \mathbb{R}^L$ as given and chooses a production plan $y_j \in \mathcal{Y}_j$ to maximize its profits:

$$\max_{y_j \in \mathcal{Y}_j} p \cdot y_j.$$

Our definition of Walrasian equilibrium extends naturally to production economies.

Definition 8. A **Walrasian equilibrium** for the production economy $\mathcal{E}$ is a vector $\left(p^*, (x_i^*)_{i \in \mathcal{I}}, \left(y_j^*\right)_{j \in \mathcal{J}}\right)$ that satisfies:

1. Firm profit maximization: for all $j \in \mathcal{J}$,

$$y_j^* \in \operatorname*{argmax}_{y_j \in \mathcal{Y}_j} p \cdot y_j,$$

2. Consumer optimization: for all consumers $i \in \mathcal{I}$,

$$x_i^* \in \operatorname*{argmax}_{x_i \in \mathcal{X}_i} u_i(x_i)$$

   subject to

$$p \cdot x_i \leq p \cdot \omega_i + \sum_{j \in \mathcal{J}} \theta_{i,j} p \cdot y_j^*,$$

3. Market-clearing: for all commodities $l \in \mathcal{L}$

$$\sum_{i \in \mathcal{I}} x_{l,i}^* = \sum_{i \in \mathcal{I}} \omega_{l,i} + \sum_{j \in \mathcal{J}} y_{l,j}^*.$$

Assumptions on Production Sets Just as we made a number of assumptions on consumer preferences and endowments, we will make several assumptions on production sets to ensure that a Walrasian equilibrium exists in a production economy. The simplest such assumption would be that Y_j is a convex and compact set for all firms $j \in \mathcal{J}$, but assuming that a production set is bounded is stronger than we need.

Assumption A5 (closed and convex): For all firms $j \in \mathcal{J}$, $\mathcal{Y}_j$ is closed and convex.

Assumption A6 (no production is feasible and free disposal): For all firms $j \in \mathcal{J}$, $0 \in \mathcal{Y}_j$, and for all $y_j \in \mathcal{Y}_j$, $\{y_j\} + \mathbb{R}^L_{-}\ \_ \subset \mathcal{Y}_j$.

These two assumptions rule out increasing returns to scale. To see why, note that if $y \in \mathcal{Y}_l$, then since $0 \in \mathcal{Y}_l$, so is αy_l for any $0 < \alpha < 1$, so it is always possible to scale down production or break it up into arbitrarily small productive units.

We will also need to make one further assumption on aggregate production to ensure that the supply side of the economy as a whole cannot produce something with nothing. We want to rule out, for example, situations where one firm can turn one pound of coffee beans into one cup of coffee, while another firm can turn one cup of coffee into two pounds of coffee beans. Define the **aggregate production set** to be the Minkowski sum of all the firms' production sets:

$$\mathcal{Y} = \sum_{j \in \mathcal{J}} \mathcal{Y}_j = \left\{ y : \text{there exist } y_1 \in \mathcal{Y}_1, \ldots, y_J \in \mathcal{Y}_J \text{ such that } y = \sum_{j \in \mathcal{J}} y_j \right\}.$$

The following assumption is sufficient to rule out the implausible situations described above.

Assumption A7 (irreversibility): $\mathcal{Y} \cap -\mathcal{Y} = \{0\}$.

It is worth spending some time thinking about why assumptions $(A6)$ and $(A7)$ rule out the situations I just described.

Welfare Theorems and Existence of Walrasian Equilibrium The definitions of feasibility and Pareto efficiency are easily extended to production economies.

Definition 9. An allocation and production plan $\left((x_i)_{i\in\mathcal{I}}, (y_j)_{j\in\mathcal{J}}\right)$ is **feasible** if

$$\sum_{i\in\mathcal{I}} x_{l,i} \leq \sum_{i\in\mathcal{I}} \omega_{l,i} + \sum_{j\in\mathcal{J}} y_{l,j} \text{ for all } l \in \mathcal{L}.$$

A feasible allocation and production plan $\left((x_i)_{i\in\mathcal{I}}, (y_j)_{j\in\mathcal{J}}\right)$ is **Pareto optimal** if there is no other feasible allocation and production plan $\left((\hat{x}_i)_{i\in\mathcal{I}}, (\hat{y}_j)_{j\in\mathcal{J}}\right)$ satisfying $u_i(\hat{x}_i) \geq u_i(x_i)$ for all i, with strict inequality for at least one i'.

We can now state the extensions of the two welfare theorems.

Theorem 9 (First Welfare Theorem). Suppose $\left(p^*, (x_i^*)_{i\in\mathcal{I}}, (y_j^*)_{j\in\mathcal{J}}\right)$ is a Walrasian equilibrium for production economy $\mathcal{E}$. Then if $(A2)$ holds, the allocation and production $\left((x_i^*)_{i\in\mathcal{I}}, (y_j^*)_{j\in\mathcal{J}}\right)$ is Pareto optimal.

The proof of the first welfare theorem for production economies is essentially the same as the proof for pure exchange economies. It is worth trying to extend each of the steps from our previous proof to allow for production. The second welfare theorem can be similarly extended.

Theorem 10 (Second Welfare Theorem). Let $\mathcal{E}$ be a production economy that satisfies $(A1)-(A6)$. Suppose $\left((x_i)_{i\in\mathcal{I}}, (y_j)_{j\in\mathcal{J}}\right)$ is Pareto optimal, and suppose $x_i >> 0$ for all $i \in \mathcal{I}$. Then there is a price vector p, ownership shares $(\theta_{i,j})_{i\in\mathcal{I},j\in\mathcal{J}}$, and endowments $(\omega_i)_{i\in\mathcal{I}}$ such that $\left(p, (x_i)_{i\in\mathcal{I}}, (y_j)_{j\in\mathcal{J}}\right)$

is a Walrasian equilibrium given these endowments and ownership shares.

The proof of the second welfare theorem again relies on the separating hyperplane theorem. Whereas the separating hyperplane in the earlier proof separated the aggregate demand set (i.e., the set of points preferred to the endowment) from the endowment, the proof in production economies requires separation between the aggregate demand set and a suitably constructed aggregate supply set (i.e., the endowment plus the set of feasible aggregate production plans). Convexity of production sets is required in order to invoke the separating hyperplane theorem.

Finally, we can also show that if we impose all the assumptions $(A1) - (A7)$, then a Walrasian equilibrium exists.

Theorem 11 (Existence of Equilibrium). Let $\mathcal{E}$ be a production economy that satisfies $(A1) - (A7)$. Then there exists a Walrasian equilibrium of $\mathcal{E}$.

Exercise 12. Consider an economy with two consumers and two commodities. Consumer 1's endowment vector is $(\lambda, 0)$ and consumer 2's is $(\mu, 0)$. Each consumer's utility is the sum of their consumption of the two commodities. Consumer 1 owns a technology for transforming commodity 1 into commodity 2. The production function is $Y = X^2$, where X is the input of commodity 1.

(a) Does this economy have a Walrasian equilibrium?

(b) What allocation would a planner choose to maximize the sum of utilities? [Be careful about second-order conditions.]

(c) What is the core of this economy?

Exercise 13 (Adapted from MWG, 16.F.2-4). In the first week, we discussed the first-order conditions for Pareto optimality in exchange economies.

This exercise asks you to extend these conditions to production economies with I consumers and J firms. Define the utility possibility set:

$$\mathcal{U} = \left\{ (u_1, \ldots, u_I) \in \mathbb{R}^I : \exists \text{ feasible } (x_i)_{i \in \mathcal{I}}, (y_j)_{j \in \mathcal{J}} \text{ with } u_i(x_i) \geq u_i \text{ for all } i \right\}.$$

Assume the production set for firm j takes the form $\mathcal{Y}_j = \left\{ y \in \mathbb{R}^L : F_j(y) \leq 0 \right\}$, where $F_j(y) = 0$ defines firm j's **transformation frontier**, and $F_j : \mathbb{R}^L \to \mathbb{R}$ is twice continuously differentiable with $F_j(0) \leq 0$ and $\nabla F_j(y_j) >> 0$ for all $y_j \in \mathbb{R}^L$.

(a) Show that if F_j is a convex function, then $\mathcal{Y}_j$ is a convex set.

(b) [Optional] Show that if, for all $i \in \mathcal{I}$, $\mathcal{X}_i$ is convex and u_i is concave, and for all $j \in \mathcal{J}$, F_j is convex, then $\mathcal{U}$ is a convex set.

(c) Suppose $\lambda \geq 0$ is a non-zero vector of Pareto weights, and consider the Pareto problem

$$\max_{u \in \mathcal{U}} \lambda \cdot u.$$

Show that the optimality conditions for an interior solution (i.e. $x_i >> 0$ for all i) for this problem satisfy

$$\frac{\partial u_i / \partial x_{l,i}}{\partial u_i / \partial x_{l',i}} = \frac{\partial u_{i'} / \partial x_{l,i'}}{\partial u_{i'} / \partial x_{l',i'}} \text{ for all } i, i', l, l' \quad (1)$$

$$\frac{\partial F_j / \partial y_{l,j}}{\partial F_j / \partial y_{l',j}} = \frac{\partial F_{j'} / \partial y_{l,j'}}{\partial F_{j'} / \partial y_{l',j'}} \text{ for all } j, j', l, l' \quad (2)$$

$$\frac{\partial u_i / \partial x_{l,i}}{\partial u_i / \partial x_{l',i}} = \frac{\partial F_j / \partial y_{l,j}}{\partial F_j / \partial y_{l',j}} \text{ for all } i, j, l, l'. \quad (3)$$

(d) Consider the aggregate problem of maximizing the production of commodity 1 subject to minimum production levels $(\bar{y}_2, \ldots, \bar{y}_L)$ for the other commodities.

$$\max_{(y_1, \ldots, y_J)} \sum_{j \in \mathcal{J}} y_{1,j}$$

subject to

$$\sum_{j \in \mathcal{J}} y_{l,j} \geq \bar{y}_l \text{ for all } l = 2, \ldots, L$$

and

$$F_j(y_j) \leq 0 \text{ for all } j = 1, \ldots, J.$$

Show that the optimality conditions for this problem satisfy (2). What does these conditions imply about how production is carried out across firms in a Pareto optimal allocation?

A Constant Returns-to-Scale Example

For a production economy, we have to specify not only consumers' preferences but also firms' production sets. A simple class of production sets that satisfy assumptions $(A5)-(A7)$ are linear production sets. Such production sets are convex cones spanned by finitely many rays.[1] There is a single firm that has access to M linear activities $a_m \in \mathcal{M}=\{1,\ldots,M\}$, and it can operate each activity at some level γ_m. Its production set $\mathcal{Y}$ is the convex hull of these activities:

$$\mathcal{Y}=\left\{y\in\mathbb{R}^L : y=\sum_{m=1}^{M}\gamma_m a_m \text{ for some } \gamma\in\mathbb{R}_+^M\right\}.$$

Assumption $(A5)$ is satisfied, and the free disposal part of assumption $(A6)$ is satisfied if the vectors

$$(-1,0,\ldots,0),(0,-1,0,\ldots,0),\ldots,(0,\ldots,0,-1)$$

[1] Let $\mathcal{X}\subset\mathbb{R}^N$ be a set that contains $\{0\}$. Take two vectors $x,y\in X$. We will say that a vector $w=\alpha x+\beta y$, where $\alpha\geq 0,\beta\geq 0$ is a **conic combination** of the vectors x and y. If the set $\mathcal{X}$ contains all conic combinations of its elements, we say that $\mathcal{X}$ is a **convex cone**. The way to think about a convex cone is to imagine a convex set $\mathcal{A}$ located some distance from the origin. The convex cone generated by the set $\mathcal{A}$ is the set of all points that lie on a ray from the origin that goes through any point in $\mathcal{A}$. If $\mathcal{A}$ is a disk, then the convex cone generated by $\mathcal{A}$ is what you would normally think of as a cone.

are all in $\mathcal{M}$.

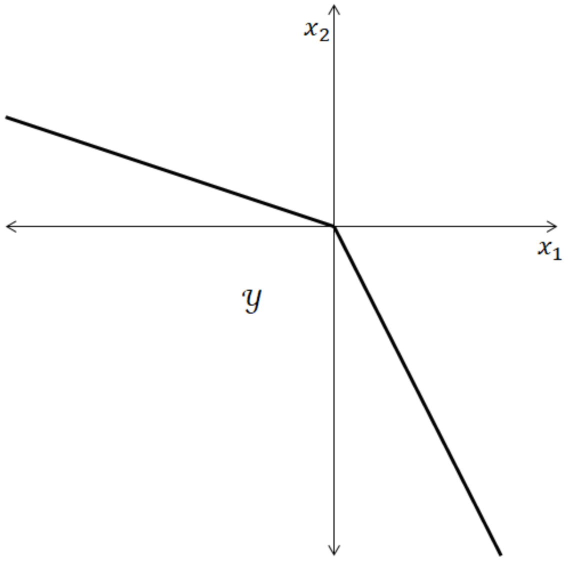


Figure 14: Linear Production Set

Figure 14 illustrates a linear production set in the special case of $M = 4$ and $L = 2$. There are two productive activities: activity 1 allows 2 units of commodity 2 to be converted into 1 unit of commodity 1. Activity 2 allows 3 units of commodity 1 to be converted into commodity 2. Activities 3 and 4 are the activities I described above that ensure that the free disposal assumption is satisfied. In this case,

$$\mathcal{M} = \{(1, -2), (-3, 1), (0, -1), (-1, 0)\}.$$

Note that the production set $\mathcal{Y}$ generated by $\mathcal{M}$ is the same production set that would be generated by activities $\{(1,-2),(-3,1)\}$ because $\frac{3}{5}(1,-2)+\frac{1}{5}(-3,1)=(0,-1)$ and $\frac{1}{5}(1,-2)+\frac{3}{5}(-3,1)=(-1,0)$.

Given a price vector p, a profit-maximizing plan exists if and only if $p\cdot a_m \leq 0$ for all $m=1,\ldots,M$. If this were not the case, the firm's potential profits would be unbounded: if $p\cdot a_m>0$ for some m, the firm could choose a sequence of production vectors $\gamma_m a_m$ with $\gamma_m\rightarrow\infty$, and its profits would increase without bound along that sequence. If $p\cdot a_m<0$ for some m, then it is clear that $\gamma_m=0$.

When production sets are convex cones, as in this example, market clearing implies that equilibrium prices are pinned down by zero-profit conditions. This specification of production sets is not as much of a special case as it first appears—it is satisfied by any constant returns to scale production technology, including the Cobb-Douglas production functions you have probably used in macroeconomics.

The Representative Firm Theorem When production sets are convex cones, firms do not really play much of a role in the economy—they earn zero profits in equilibrium, and it is actually irrelevant whether there is a single firm that possesses the entire set of activities $\mathcal{M}$ or a collection of M firms that each possess only a activity a_m (plus the free disposal activities). The result that it is without loss of generality to focus on a single firm that possesses the sum of firms' production technologies is actually a much more general result than this example illustrates, as the following theorem (Acemoglu, 2009) highlights.

Theorem 12 (Representative Firm Theorem). Consider a competitive production economy with L commodities and J firms, each with production-possibilities set $\mathcal{Y}_j \subset \mathbb{R}^L$. Let $p \in \mathbb{R}_+^L$ be the price vector in this economy and denote the set of profit-maximizing net supplies of firm $j \in \mathcal{J}$ by $\tilde{\mathcal{Y}}_j(p) \subset \mathcal{Y}_j$ (so that for any $\tilde{y}_j \in \tilde{\mathcal{Y}}_j(p)$, we have $p \cdot \tilde{y}_j \geq p \cdot y_j$ for all $y_j \in Y_j$). Then there exists a representative firm with production possibilities set $\mathcal{Y} \subset \mathbb{R}^L$ and a set of profit-maximizing net supplies $\tilde{\mathcal{Y}}(p)$ such that for any $p \in \mathbb{R}_+^L$, $\tilde{y} \in \tilde{\mathcal{Y}}(p)$ if and only if $\tilde{y} = \sum_{j \in \mathcal{J}} \tilde{y}_j$ for some $\tilde{y}_j \in \tilde{\mathcal{Y}}_j(p)$ for each $j \in \mathcal{J}$.

This theorem shows us that we can "aggregate" the production side of the economy. Exercise 13 asks you to prove this result. The rough idea of the theorem is that the analog of the conditions required for aggregation of consumer preferences is always satisfied for firms.

Exercise 14. This exercise asks you to prove the representative firm theorem.

(a) Fix p and construct $\tilde{y} = \sum_{j \in \mathcal{J}} \tilde{y}_j$ for some $\tilde{y}_j \in \tilde{\mathcal{Y}}_j(p)$ for each $j \in \mathcal{J}$. Prove that we must have $\tilde{y} \in \tilde{\mathcal{Y}}(p)$.

(b) Let $\tilde{y} \in \tilde{\mathcal{Y}}(p)$ be a profit-maximizing choice for the representative firm. Show that if $\tilde{y} = \sum_{j \in \mathcal{J}} y_j$ for some $y_j \in \mathcal{Y}_j$ for each $j \in \mathcal{J}$, then $y_j \in \tilde{\mathcal{Y}}_j(p)$ for each $j \in \mathcal{J}$.

3.2 Uncertainty and Time in General Equilibrium

Another, and perhaps the most important, extension to the general equilibrium framework is to allow for both time and uncertainty. Introducing time into the framework is straightforward: we can think of a consumption good today as a different commodity than a consumption good tomorrow. Adding uncertainty turns out also to be straightforward due to an important modeling device of Arrow (1953/1964): states of the world. A state of the world is a complete description of a date-event. Everyone agrees on the set of possible states and what state of the world is realized, although they need not necessarily agree on the probabilities of those states occurring. This way of thinking about uncertainty makes it very easy to extend the general equilibrium framework. In fact, Debreu's (1959) *Theory of Value* devotes only a short chapter to general equilibrium under uncertainty, and in some sense the first paragraph of that chapter tells us the main idea.

> "The analysis is extended in this chapter to the case where uncertain events determine the consumption sets, the production sets, and the resources of the economy. A contract for the transfer of a commodity now specifies, in addition to its physical properties, its location and its date, an event on the occurrence of which the transfer is conditional. This new definition of a commodity allows one to obtain a theory of uncertainty free from any proba-

> bility concept and formally identical with the theory of certainty developed in the preceding chapters." (Debreu, 1959)

Arrow-Debreu Model

We will consider a parsimonious model of GE with uncertainty, although the framework can accommodate much more general specifications. Suppose there are I consumers $i \in \mathcal{I}$, L consumption goods $l \in \mathcal{L}$, and two periods, $t \in \{0, 1\}$. There are S possible states of the world that can occur at $t = 1$, $s \in \mathcal{S} = \{1, \ldots, S\}$. A consumption bundle for consumer i is an $x_i = (x_{0,i}, x_{1,i}, \ldots, x_{S,i})$, where $x_{0,i} = (x_{1,0,i}, \ldots, x_{L,0,i})$ is consumer i's consumption of the L goods at $t = 0$, and $x_{s,i} = (x_{1,s,i}, \ldots, x_{L,s,i})$ is her consumption of the L goods at $t = 1$ and in state s. Her consumption set is $\mathcal{X}_i = \mathbb{R}_+^{L(S+1)}$, and her preferences are given by her utility function $U_i : \mathbb{R}_+^{L(S+1)} \to \mathbb{R}$. She has endowment $\omega_i = (\omega_{0,i}, \omega_{1,i}, \ldots, \omega_{S,i})$, where $\omega_{0,i} = (\omega_{1,0,i}, \ldots, \omega_{L,0,i})$ is her endowment of the L goods at $t = 0$, and $\omega_{s,i} = (\omega_{1,s,i}, \ldots, \omega_{L,s,i})$ is her endowment of the L goods at $t = 1$ in state s.

Since there is uncertainty, we also have to specify consumers' beliefs. Suppose, at $t = 0$, consumer i believes that state $s \in \mathcal{S}$ will occur with probability $\pi_{s,i} \geq 0$, where $\sum_{s \in \mathcal{S}} \pi_{s,i} = 1$ for all i. Typically, we will think about consumers having the same beliefs, so that $\pi_{s,i} = \pi_s$ for all i, but the framework allows for subjective beliefs. We will also typically assume that consumers are expected utility maximizers with additively separable time preferences, so that we can write

$$U_i\left(x_i\right) = u_{0,i}\left(x_{0,i}\right) + \sum_{s\in\mathcal{S}} \pi_{s,i} u_{s,i}\left(x_{s,i}\right),$$

with $u_{0,i}$ and each of the $u_{s,i}$ functions concave.

In line with Debreu's description, we will think of each consumption good in each state of the world as being a separate commodity. To specify prices, therefore, we will have to specify $p = (p_0, p_1, \ldots, p_S)$, where $p_0 = (p_{1,0}, \ldots, p_{L,0}) \in \mathbb{R}^L$ is the price vector at $t = 0$, and $p_s = (p_{1,s}, \ldots, p_{L,s}) \in \mathbb{R}^L$ is the price vector at $t = 1$ in state $s \in \mathcal{S}$. That is, for a price $p_{l,s}$, consumers can buy and sell consumption of good l in state s.

There are three important assumptions that allow us to make use of all of the results we have derived so far in this course. The first assumption is that all trade occurs at time $t = 0$. So, at time $t = 0$, consumers buy and sell $t = 0$ commodities, and they also buy future claims to each commodity in each state of the world, and there is no opportunity for them to buy or sell at $t = 1$. The second important assumption is that the trading contracts that each consumer "writes" at $t = 0$ over $t = 1$ consumer are faithfully executed at $t = 1$. In the background, we are implicitly assuming the existence of an infallible third-party court system that perfectly compels consumers to execute their $t = 0$ contracts. This assumption, in turn, means that the third party can costlessly verify what state of the world was actually realized at $t = 1$. The third important assumption is that there is a market for each of the $L(S+1)$ state-contingent commodities.

Given prices p, consumer i solves

$$\max_{x_i \in \mathbb{R}_+^{L(S+1)}} U_i(x_i) \text{ s.t. } x_i \in \mathcal{B}_i(p),$$

where consumer i's budget set is given by

$$\mathcal{B}_i(p) = \left\{ x_i : p_0 \cdot x_{0,i} + \sum_{s \in \mathcal{S}} p_s \cdot x_{s,i} \leq p_0 \cdot \omega_{0,i} + \sum_{s \in \mathcal{S}} p_s \cdot \omega_{s,i} \right\}.$$

We will denote her Marshallian demand correspondence by $x_i(p, p \cdot x_i)$. A **pure-exchange economy with uncertainty** is therefore summarized by $\mathcal{E} = (u_i, \omega_i)_{i \in \mathcal{I}}$.

We are now in a position to define a Walrasian equilibrium in this context. For historical reasons, Walrasian equilibria in this model are referred to as Arrow-Debreu equilibria.

Definition 10. An **Arrow-Debreu equilibrium** for pure-exchange economy with uncertainty $\mathcal{E}$ is a vector $\left(p^*, (x_i^*)_{i \in \mathcal{I}}\right)$ that satisfies:

1. Consumer optimization: for all consumers $i \in \mathcal{I}$,

$$x_i^* \in \operatorname*{argmax}_{x_i \in \mathcal{B}_i(p^*)} U_i(x_i),$$

2. Market-clearing: for all commodities $l \in \mathcal{L}$ and all $s \in \{0, 1, \ldots, S\}$,

$$\sum_{i \in \mathcal{I}} x^*_{l,s,i} = \sum_{i \in \mathcal{I}} \omega_{l,s,i}.$$

This model is an elegant way of incorporating time and uncertainty into the basic framework because it allows us to apply all the results we have developed so far. For example, if $(A1)-(A4)$ hold, then a Walrasian equilibrium exists, and the welfare theorems hold.

The model does have some issues, though. One natural concern is that it seems unrealistic to think of all trading over future state-contingent commodities taking place at the beginning of time. Instead, we might expect that there would be different financial securities that are traded at potentially different times and these securities pay out when certain events occur. For example, car insurance pays out when your car is stolen, stocks pay dividends when a company is doing well, and so on.

Exercise 15. Consider a two date exchange economy with consumption at dates 0 and 1. There is a single consumer, one consumption good at each date, and there are S states of the world (realized at date 1).

The consumer's utility function is

$$U = u(x_0) + \delta \sum_{s=1}^{S} \pi_s u(x_s),$$

where x_0 is date 0 consumption, x_s is date 1 consumption in state s, u is "well-behaved," and $\delta \in (0,1)$. The consumer has initial endowment $(\omega_0, \omega_1, \ldots, \omega_S) \in \mathbb{R}^{S+1}_+$.

Write down the Arrow-Debreu equilibrium for this economy (normalize the price of date-0 consumption to be 1). Interpret the Arrow-Debreu relative prices: what factors determine whether they are high or low?

Sequential Trade and Arrow Equilibrium

Arrow later reformulated the model to allow for sequential trade in the following way. As before, there are two dates, $t \in \{0, 1\}$, and at date $t = 1$, a state of the world $s \in \mathcal{S}$ is realized. Suppose that consumption occurs only at $t = 1$, so that all consumers are endowed with $\omega_{l,0,i} = 0$ for all $l \in \mathcal{L}$, and $x_{0,i} \in \{0\}$ for all i. Moreover, suppose consumer i is an expected utility maximizer, so that $U_i(x_i) = \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i}(x_{s,i})$.

At date $t = 0$, consumers cannot directly trade all $L \cdot S$ state-contingent commodities. They can, however, trade securities that pay off different amounts in different states at $t = 1$. In particular, they can trade S different Arrow securities, where at $t = 1$, **Arrow security** s pays \$1 if state s is realized, and it pays 0 otherwise. Each consumer is endowed with 0 of each Arrow security, but they can have positive or negative holdings of them after trade occurs at $t = 0$. Denote by $z_i = (z_{1,i}, \ldots, z_{S,i})$ consumer i's holdings of the S Arrow securities, and denote the price vector for the Arrow securities by $q = (q_1, \ldots, q_S)$. To anticipate how we will think of more general securities in the next section, denote the **dividends vector for security** k by $r_k = (r_{1,k}, \ldots, r_{S,k}) \in \mathbb{R}_+^S$, where $r_{s,k}$ is the amount that security k pays in state s. The dividends vector for Arrow security 1 is therefore $r_1^A \equiv (1, 0, \ldots, 0)$, and for Arrow security k is $r_k^A \equiv (0, \ldots, 0, 1, 0, \ldots, 0)$, where the kth element is 1 and all others are 0.

At $t = 1$, the state of the world s is realized, and then markets for each of the L goods open, consumers trade at prices $p_s = (p_{s,1}, \ldots, p_{s,L})$, and then they consume.

Under this specification, given Arrow security prices q and goods prices p, consumer i solves the following problem:

$$\max_{(z_i,x_i)} \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i}(x_{s,i}) \text{ s.t. } (z_i, x_i) \in \mathcal{B}_i(q,p),$$

where her budget set is now given by

$$\mathcal{B}_i(q,p) = \{(z_i, x_i) : q \cdot z_i \leq 0, p_s \cdot x_{s,i} \leq p_s \cdot \omega_{s,i} + z_{s,i} \text{ for all } s \in \mathcal{S}\}.$$

The first inequality in the definition of the budget set reflects the assumption that the consumer is not endowed with any Arrow securities, so that the net value of the Arrow securities she holds after $t = 0$ trade has to be nonpositive. The second set of inequalities reflects her budget set at $t = 1$ in state s. Her wealth in state s is the sum of the wealth from her endowment, $p_s \cdot \omega_{s,i}$, and the wealth she obtains from her Arrow securities, $z_{s,i}$, which can be positive or negative. Note that, since we are assuming that $x_i \in \mathbb{R}_+^{L \cdot S}$, we are implicitly imposing the constraint that each consumer has nonnegative wealth at $t = 1$: $z_{s,i} \geq -p_s \cdot \omega_{s,i}$.

The idea behind this alternative setup of the model is that consumers will trade multiple times, and their wealth each time they trade is determined by their endowment in the "spot market" as well as how much they loaned and borrowed. Consumers correctly anticipate spot-market prices in each state at $t = 0$, even though they cannot trade in those markets until $t = 1$, and they buy and sell Arrow securities to transfer their wealth from one state to the next so they can buy the commodities they would like to buy in those states. We will refer to the economy as a **sequential-exchange economy with a complete set of Arrow securities** and denote it by $\mathcal{E}^{SE} = \left((u_i, \omega_i)_{i \in \mathcal{I}}, \left(r_k^A \right)_{k \in \mathcal{S}} \right)$, where r_k^A is the returns vector for the kth Arrow security.

We can now define our notion of Walrasian equilibrium in this setting.

Definition 11. An **Arrow equilibrium** for a sequential-exchange economy with a complete set of Arrow securities, $\mathcal{E}^{SE}$, is a vector $\left(q^*, p^*, (z_i^*, x_i^*)_{i \in \mathcal{I}} \right)$ of Arrow security prices and state-contingent consumption-good prices and Arrow security positions and consumption bundles for each consumer $i \in \mathcal{I}$ that satisfies:

1. Consumer optimization: for all consumers $i \in \mathcal{I}$,

$$(z_i^*, x_i^*) \in \underset{(z_i, x_i) \in \mathcal{B}_i(q,p)}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i} \left(x_{s,i} \right),$$

2. Market-clearing: $\sum_{i \in \mathcal{I}} z_i^* = 0$ and, for all commodities $l \in \mathcal{L}$ and all $s \in \mathcal{S}$,

$$\sum_{i \in \mathcal{I}} x_{l,s,i}^* = \sum_{i \in \mathcal{I}} \omega_{l,s,i}.$$

Given this definition of equilibrium, we can now describe the main result of this section, which links the set of allocations that can arise in an Arrow-Debreu equilibrium in a pure-exchange economy with uncertainty, $\mathcal{E} = (u_i, \omega_i)_{i \in \mathcal{I}}$ to the set of allocations that can arise in an Arrow equilibrium in a sequential-exchange economy with a complete set of Arrow securities, $\mathcal{E}^{SE} = \left((u_i, \omega_i)_{i \in \mathcal{I}}, \left(r_k^A \right)_{k \in \mathcal{S}} \right)$.

Theorem 13 (Equivalence of Arrow and Arrow-Debreu equilibrium). Given economies $\mathcal{E}$ and $\mathcal{E}^{SE}$ with the same consumer preferences and endowments, $(x_i^*)_{i \in \mathcal{I}}$ is an Arrow-Debreu equilibrium allocation for $\mathcal{E}$ if and only if, for some $(z_i^*)_{i \in \mathcal{I}}$, $(z_i^*, x_i^*)_{i \in \mathcal{I}}$ is an Arrow equilibrium allocation for $\mathcal{E}^{SE}$.

Proof of Theorem 13. Take an Arrow equilibrium $\left(q^*, p^*, (z_i^*, x_i^*)_{i \in \mathcal{I}} \right)$ for economy $\mathcal{E}^{SE}$. By monotonicity of preferences, we will have that for each consumer $i \in \mathcal{I}$, $q^* \cdot z_i^* (q^*, p^*) = 0$ and $z_{s,i}^* (q^*, p^*) = p_s^* \cdot x_{s,i}^* (q^*, p^*) - p_s^* \cdot \omega_{s,i}$. We can combine these two equations to get

$$\sum_{s \in \mathcal{S}} q_s^* \left(p_s^* \cdot x_{s,i}^* - p_s^* \cdot \omega_{s,i} \right) = 0.$$

The consumption bundle $x_{s,i}^*$ therefore solves the problem:

$$\max_{x_i} \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i} \left(x_{s,i} \right)$$

subject to

$$\sum_{s \in \mathcal{S}} \left(q_s^* p_s^* \right) \cdot x_{s,i} \leq \sum_{s \in \mathcal{S}} \left(q_s^* p_s^* \right) \cdot \omega_{s,i}.$$

Define state-dependent prices $\hat{p}_s^* \in \mathbb{R}^L$ for $s \in \mathcal{S}$ with $\hat{p}_s^* \equiv q_s^* p_s^*$. Then, $\left(\hat{p}^*, (x_i^*)_{i \in \mathcal{I}}\right)$ is an Arrow-Debreu equilibrium for economy $\mathcal{E}$.

Going the other direction, suppose $\left(p^*, (x_i^*)_{i \in \mathcal{I}}\right)$ is an Arrow-Debreu equilibrium for $\mathcal{E}$. Then $\left(q^*, \hat{p}^*, (z_i^*, x_i^*)_{i \in \mathcal{I}}\right)$, where $q_s^* = 1$, $\hat{p}^* = p^*$, and

$$z_{s,i}^* = p_s^* \cdot x_{s,i}^* - p_s^* \cdot \omega_{s,i}$$

is an Arrow equilibrium for economy $\mathcal{E}^{SE}$.■

This theorem establishes an equivalence between the notion of Arrow-Debreu equilibrium in which trade in all $L \cdot S$ markets occurs ex ante and the notion of Arrow equilibrium, in which trade occurs in only S markets at $t = 0$ and in L markets at $t = 1$. One disadvantage of the notion of Arrow equilibrium is that even though trading seems less complicated, in a sense, consumers still must form consistent expectations about what the equilibrium goods prices will be at $t = 1$ when they are trading securities at $t = 0$. That said, one of the big advantages of the sequential exchange framework is that it allows us to investigate what happens when the $t = 0$ securities market does not have a complete set of Arrow securities. That is, what happens if markets are *incomplete*? We will turn to this question now.

Exercise 16. There are two farmers, named Octavia and Seema, who can trade only with each other. In years when there is no flood, both farms yield 10 units of corn; in years when there is a flood, Octavia's farm yields 10 units of corn, and Seema's farm yields 5 units of corn. The probability of a flood is given by $\pi = 1/2$, which is common knowledge to the farmers. The farmers have identical utility functions given by $u(x) = \ln(x)$, where x is the units of corn consumed.

(a) Suppose that Octavia and Seema set up an exchange market to securitize corn at the beginning of the year (before knowing the realization of the state of the world). Compute the equilibrium prices and allocations.

Suppose that Seema has the option of building a greenhouse at a cost before realizing the state of the world. If she builds a greenhouse, Seema's farm will produce 10 units of corn in all states of the world.

(b) Using the equilibrium results computed above, how much would each farmer be willing to pay for the greenhouse? Assume that each considers paying for the greenhouse entirely by herself. In this context, should we consider the possibility of "negative" willingness-to-pay? That is, might one farmer be willing to pay the other not to build the greenhouse?

(c) Would your above answer change if Octavia and Seema were unable to trade ex post (after the state of the world is realized)? If so, how? Would your answer change if they were unable to trade ex ante (there is no exchange market to securitize corn at the beginning of the year)?

Incomplete Markets

When we talked about sequential exchange economies in the previous section, we assumed that there was a complete set of Arrow securities that could be traded at $t = 0$. One important implication of this assumption that we did not emphasize is that it allowed consumers to insure themselves against the state of the world by transferring wealth from states in which their marginal utility of income is low (either because they do not especially value consumption in such states or because their endowment in such states is high) to states in which their marginal utility of income is high. In an Arrow equilibrium, the resulting risk-sharing is efficient, since the first welfare theorem applies in that setting.[2] In contrast, when markets are *incomplete*, risk sharing in the economy will generally be inefficient. This will imply that Walrasian equilibrium allocations in such economies are not Pareto optimal.

Suppose there are K **securities** $k \in \mathcal{K} = \{1, \ldots, K\}$, where security k has **dividends vector** $r_k = (r_{1,k}, \ldots, r_{S,k}) \in \mathbb{R}_+^S$. We can think of each security as a share in a company that pays out dividends $r_{s,k}$ in state s. If consumer i owns **portfolio** $z_i = (z_{1,i}, \ldots, z_{K,i})$, then in state s, her wealth will be $p_s \cdot \omega_{s,i} + \sum_{k \in \mathcal{K}} z_{k,i} r_{s,k}$. If we denote the **dividends matrix** $R = \left(r_1^T, \ldots, r_K^T\right)$, where r_k^T is the transpose of r_k, then we will say that the securities market is **incomplete** if $rank\,(R) < S$. Otherwise, we will say that the securities market is **complete**. If there is a complete set of Arrow securities, then R is the $S \times S$ identity matrix, and the securities market is complete.

As in the previous subsection, given security prices $q = (q_1, \ldots, q_K)$ and goods prices p, consumer $i \in \mathcal{I}$ solves the following problem:

$$\max_{(z_i, x_i)} \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i}\left(x_{s,i}\right) \text{ s.t. } \left(z_i, x_i\right) \in \mathcal{B}_i\left(q, p\right),$$

where her budget set is now given by

$$\mathcal{B}_i\left(q, p\right) = \left\{ \left(z_i, x_i\right) : q \cdot z_i \leq 0, p_s \cdot x_{s,i} \leq p_s \cdot \omega_{s,i} + \sum_{k \in \mathcal{K}} z_{k,i} r_{s,k} \text{ for all } s \in \mathcal{S} \right\}.$$

[2] More precisely, define $v_{s,i}\left(z_{s,i}\right) = \max_{x_{s,i}} u_{s,i}\left(x_{s,i}\right)$ subject to $p_s \cdot x_{s,i} \leq p_s \cdot \omega_{s,i} + z_{s,i}$ to be consumer i's indirect utility in state s when she has $z_{s,i}$ units of Arrow security s. If we assume $v_{s,i}$ is concave and differentiable, then Pareto optimality ensures that for all i, i', $\frac{\pi_{s,i} \partial v_{s,i} / \partial z_{s,i}}{\pi_{s,i'} \partial v_{s,i'} / \partial z_{s,i'}} = \frac{\lambda_i}{\lambda_{i'}}$ for all s. That is, the ratio of marginal utilities of income are equalized across sttaes for any two consumers. By the first welfare theorem, any Arrow equilibrium allocation satisfies these properties.

As before, consumers maximize their expected utility subject to a $t = 0$ budget constraint and a $t = 1$ budget constraint for each state $s \in \mathcal{S}$. The first inequality in the definition of the budget set again reflects the assumption that the consumer is not endowed with any securities.

A **sequential-exchange economy with securities** $\mathcal{K}$ is summarized by a vector $\mathcal{E}^{SE} = \left((u_i, \omega_i)_{i \in \mathcal{I}}, R \right)$. We can now define our notion of Walrasian equilibrium for such an economy.

Definition 12. An **incomplete-markets equilibrium (or Radner equilibrium)** for a sequential-exchange economy with securities $\mathcal{K}$ is a vector $\left(q^*, p^*, (z_i^*, x_i^*)_{i \in \mathcal{I}} \right)$ of security prices and state-contingent consumption-good prices and security positions and consumption bundles for each consumer $i \in \mathcal{I}$ that satisfies:

1. Consumer optimization: for all consumers $i \in \mathcal{I}$,

$$(z_i^*, x_i^*) \in \underset{(z_i, x_i) \in \mathcal{B}_i(q,p)}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i} \left(x_{s,i} \right),$$

2. Market-clearing: $\sum_{i \in \mathcal{I}} z_{k,i}^* = 0$ for all $k \in \mathcal{K}$, and, for all commodities $l \in \mathcal{L}$ and all $s \in \mathcal{S}$,

$$\sum_{i \in \mathcal{I}} x_{l,s,i}^* = \sum_{i \in \mathcal{I}} \omega_{l,s,i}.$$

In general, when markets are incomplete, a Radner equilibrium need not exist, and even if it does exist, the resulting allocation will typically not be Pareto optimal. If, however, $L = 1$, so there is only a single consumption good, then a Radner equilibrium exists, and it does have some optimality properties.

For $L = 1$, Diamond (1967) showed that a Radner equilibrium exists by showing that the consumer optimization problem boils down to a more familiar problem. In particular, at any solution to consumer i's problem, we must have $q \cdot z_i = 0$ and $x_{s,i} = \omega_{s,i} + \sum_{k \in \mathcal{K}} z_{k,i} r_{s,k}$. We can substitute this second constraint into the consumer's problem, which then becomes

$$\max_{z_i} \underbrace{\sum_{s \in \mathcal{S}} \pi_{s,i} u_{s,i} \left(\omega_{s,i} + \sum_{k \in \mathcal{K}} z_{k,i} r_{s,k} \right)}_{\equiv \tilde{u}_i(z_i)}$$

subject to $q \cdot z_i \leq 0$. Diamond pointed out that such an economy is equivalent, in some sense, to an economy in which consumer preferences are given by $\tilde{u}_i (z_i)$ and there are K "commodities"—one corresponding to each of the securities. So as long as $\tilde{u}_i$ satisfies $(A1) - (A3)$, then a WE exists. The interior endowments assumption $(A4)$ is not necessary for the existence result because consumers are allowed to "consume" negative quantities of z_i.

Such equilibria need not yield Pareto-optimal allocations. To see why, consider an example in which $L = 1, K = 1$, and $S = 2$. The security pays 1 in each state of the world. There are two consumers with endowments $\omega_1 = (2, 1)$ and $\omega_2 = (1, 2)$, so that consumer 1 is endowed with one unit of the consumption good in state 1 and 2 units in state 2. Both consumers have identical preferences given by

$$u_i\left(x_{1,i}, x_{2,i}\right) = \frac{1}{2}\log x_{1,i} + \frac{1}{2}\log x_{2,i}.$$

As an exercise, it is worth verifying that there is a unique Radner equilibrium of this economy. In this equilibrium, there will be no trade in the security at $t = 0$, and consumers will consume their endowments at $t = 1$. This allocation is Pareto dominated by the feasible allocation $x_1 = x_2 = (3/2, 3/2)$.

The first welfare theorem fails in this situation because the set of existing securities does not allow the consumers to insure themselves against states in which they will have a low endowment. Nevertheless, there is still a sense in which the resulting allocation exhausts the gains from trade and is therefore what we refer to as constrained efficient.

Definition 13. Given endowments $(\omega_i)_{i\in\mathcal{I}}$ and securities $\mathcal{K}$, an allocation $(x_i)_{i\in\mathcal{I}}$ is **constrained efficient** if $\sum_{i\in\mathcal{I}} x_i \leq \sum_{i\in\mathcal{I}} \omega_i$, and for all i, there exists $z_i \in \mathbb{R}^K$ such that $x_i = \omega_i + Rz_i$, and there exists no alternative allocation $(\hat{x}_i)_{i\in\mathcal{I}}$ that Pareto dominates $(x_i)_{i\in\mathcal{I}}$ and also satisfies $\sum_{i\in\mathcal{I}} \hat{x}_i \leq \sum_{i\in\mathcal{I}} \omega_i$ and $\hat{x}_i = \omega_i + Rz_i$ for some $z_i \in \mathbb{R}^K$ for all i.

When there is only a single consumption good and consumers have monotone preferences, Radner equilibrium allocations are always constrained efficient, as the following theorem illustrates.

Theorem 14. If $\mathcal{E}^{SE}$ has $L = 1$ and satisfies $(A2)$, then if $\left(q^*, p^*, (z_i^*, x_i^*)_{i\in\mathcal{I}}\right)$ is a Radner equilibrium, $(x_i^*)_{i\in\mathcal{I}}$ is constrained efficient.

We will conclude this section with a few comments on the generality of this theorem. As Hart (1975) illustrated, when $L = 2$, there may exist Radner equilibria that are not constrained efficient (see, for instance, MWG Example 19.F.2). Also, when markets are incomplete, weird things can happen. For example, adding another security that is linearly independent of existing securities can actually make all consumers strictly worse off (see, for instance, MWG Exercise 19.F.3). Finally, when markets are incomplete, Geanakoplos and Polemarchakis (1986) show that it is generically true that a social planner can improve efficiency by introducing a small tax or subsidy. This is an illustration of the "general theory of the second-best" (Lipsey and Lancaster, 1956): when there is an unresolvable market failure (market incompleteness, in this case), it is generically the case that there exists a further distortion that a social planner could conceivably put in place that leads to a more efficient allocation.

Part II

The Visible Hand

Chapter 4

Contract Theory

One of the important issues that we touched on briefly in our discussion of general equilibrium theory is the idea of market incompleteness and its consequences. When markets are incomplete—either in the sense that we talked about last time or in the sense that consumption involves unpriceable externalities—equilibrium allocations may not be constrained-efficient, opening up scope for some sort of third-party intervention. It may be government intervention via a system of taxation or rules, or it may be private intervention by an entrepreneur who sets up governance institutions. We were able to make some high-level claims last time about what happens when there are these "market failures," but without imposing more structure on the problem, it is difficult to make specific claims about how they should be managed.

For the last three weeks of the class, we will zoom in and study micro situations in which it could be said that markets are incomplete. We will focus on what is referred to as the Principal–Agent problem in which there are two players, a Principal P and an Agent A. The Principal needs the Agent to do something that she cannot do herself, so she hires the Agent and writes a contract that governs how the Agent will be paid. We can think of the Principal being an employer and the Agent an employee, where the Principal lacks the time or expertise to engage in production. We can think of the Principal being a patient and the Agent a doctor, where the doctor takes some actions that the patient does not know or understand. We can think of the Principal being a client and the Agent being a lawyer acting on the client's behalf. And so on.

When equilibrium outcomes arising from the Principal–Agent interaction are Pareto inefficient, we will say that there is a *moral hazard problem*, which is a term that originated in the insurance industry to describe situations in which someone increases their exposure to risk in response to buying insurance. Fundamentally, the moral hazard problem is a just an externality problem. Now, when we make a claim like "there are externalities, so outcomes will be inefficient" it is important to have in mind that whether or not externalities "matter" in the sense that they lead to Pareto inefficient equilibrium outcomes depends critically on the set of instruments parties have for managing those externalities: it depends on the contracting space. Over the next couple lectures, we will look at several different sources of *contractual frictions* that prevent the Principal and Agent from writing contracts with each other that result in Pareto optimal outcomes.

The first situation we will look at will occur when individual actions chosen by the Agent are not observed by the Principal but determine the distribution of a verifiable performance measure that can be written into a contract. The Agent may be more risk-averse than the Principal, so writing a high-powered contract on that noisy performance measure transfers risk onto the Agent and therefore leads to an inefficient allocation of risk between the two parties. As a result, there is a trade-off between incentive provision (and therefore what the Agent chooses to do) and inefficient risk allocation. This is the celebrated *risk–incentives trade-off*.

The second contracting friction that might arise is that an Agent is either liquidity-constrained or is subject to a limited-liability constraint. As a result, the Principal is unable to extract all the surplus the Agent generates and must therefore provide the Agent with *incentive rents* in order to motivate him. That is, offering the Agent a higher-powered contract induces him to work harder and therefore increases the total size of the pie, but it also leaves the Agent with a larger share of that pie. The Principal then, in choosing a contract, chooses one that trades off the creation of surplus with her ability to extract that surplus. This is the *motivation–rent extraction trade-off*.

A third contracting friction that might arise is that the Principal's objective simply cannot be written into a formal contract. Instead, the Principal has to rely on imperfectly aligned performance measures. Increasing the strength of a formal contract that is based on imperfectly aligned performance measures may motivate the Agent to work hard toward the Principal's objectives, but it may also motivate him to work hard toward objectives that either hurt the Principal or at least do not help her. This is known as the *multi-task problem* (Holmström and Milgrom, 1991), and failure to account for the effects of using distorted performance measures is sometimes referred to as *the folly of rewarding A while hoping for B* (Kerr, 1975).

Finally, there may be multiple Agents who work together to produce something for the Principal. Their individual contributions may not be observable, so contracts may only be able to be written on the final output. This inability to distinguish individual contributions is what is referred to as the *moral hazard in teams problem* (Holmström, 1982).

All of these sources of contractual frictions lead to similar results—under the optimal contract, the Agent (or Agents) chooses an action that is not jointly optimal from his and the Principal's perspective. But in different applied settings, different assumptions regarding what is contractible and what is not are more or less plausible. As a result, it is useful to master at least elementary versions of models capturing these four sources of frictions, so that you are well-equipped to use them as building blocks.

4.1 The Risk-Incentives Trade-off

I will begin with a pretty general description of the standard principal-agent model, but I will shortly afterwards specialize the model quite a bit in order to focus on a single point—the risk–incentives trade-off.

The Model There is a risk-neutral Principal (P) and a risk-averse Agent (A). The Agent chooses an **effort level** $e \in \mathcal{E} \subset \mathbb{R}_+$ and incurs a cost of $c(e)$, where $c : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing and strictly convex. If $\mathcal{E}$ is an interval, we will say that **effort is continuous**, and if $\mathcal{E}$ consists of a finite number of points, we will say that **effort is discrete**. We will assume $0 \in \mathcal{E}$, and $c(0) = 0$. The effort level affects the distribution over **output** $y \in \mathcal{Y}$, with y distributed according to CDF $F(\cdot|\, e)$. This output can be sold on the product market at price p, and the revenues py accrue to the Principal.

The Principal does not have any direct control over the Agent, but what she can do is write a contract that influences what the Agent will do. In particular, she can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{Y} \times \mathcal{E} \to \mathbb{R}\}$, where $\mathcal{W}$ is the **contracting space**. The contract determines a transfer $w(y, e)$ that she is compelled to pay the Agent if output y is realized, and he chose effort e. If $\mathcal{W}$ does not allow for functions that depend directly on effort, we will say that **effort is noncontractible**, and abusing notation slightly, we will write the contractual payment the Principal is compelled to pay the Agent if output y is realized as $w(y, e) = w(y)$ for all $e \in \mathcal{E}$. We will be assuming throughout that effort is noncontractible, but I wanted to highlight that it is a real restriction on the contracting space, and it is one that we will impose as a primitive of the model.

The Agent can decline to work for the Principal and reject her contract, pursuing his outside option instead. This outside option provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the Agent accepts the contract, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned}\Pi(w,e) &= \int_{y\in\mathcal{Y}} (py - w(y))\, dF(y|\, e) = E_y\left[py - w|\, e\right] \\ U(w,e) &= \int_{y\in\mathcal{Y}} u(w(y) - c(e))\, dF(y|\, e) = E_y\left[u(w - c(e))|\, e\right],\end{aligned}$$

where u is increasing and weakly concave.

We have described the players, what they can do, and what their preferences are. We still need to describe the timing of the game that the players play, as well as the solution concept. Explicitly describing the timing of the model is essential to remove any ambiguity about what players know when they make their decisions. In this model, the timing of the game is:

1. P offers A a contract $w \in \mathcal{W}$. w is commonly observed.

2. A accepts the contract ($d = 1$) or rejects it ($d = 0$), in which case he receives $\bar{u}$, and the game ends. d is commonly observed.

3. If A accepts the contract, A chooses effort level e and incurs cost $c(e)$. e is privately observed by A.

4. Output y is drawn from distribution with CDF $F(\cdot|\, e)$. y is commonly observed.

5. P pays A an amount $w(y)$. The payment is commonly observed.

A couple remarks are in order at this point. First, behind the scenes, there is an implicit assumption that there is a third-party contract enforcer (a judge or arbitrator) who can costlessly detect when agreements have been broken and costlessly exact harsh punishments on the offender.

Second, much of the literature assumes that the Agent's effort level is privately observed by the Agent and therefore refers to this model as the "hidden action" model. Ultimately, though, the underlying source of the moral-hazard problem is that contracts cannot be conditioned on relevant variables, not that the relevant variables are unobserved by the Principal. Many papers assume effort is unobservable to justify it being noncontractible. While this is a compelling justification, in our framework, the contracting space itself is a primitive of the model. Later in the course, we will talk a bit about the microfoundations for different assumptions on the contracting space.

Finally, let us describe the solution concept. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an **acceptance decision** $d^* : \mathcal{W} \to \{0,1\}$, and an **effort choice** $e^* : \mathcal{W} \times \{0,1\} \to \mathcal{E}$ such that, given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract **induces** effort e^*.

First-Best Benchmark If we want to talk about the inefficiencies that arise in equilibrium in this model, it will be useful first to establish a benchmark against which to compare outcomes. In this model, a **feasible outcome** is a distribution over payments from the Principal to the Agent as well as an effort level $e \in \mathcal{E}$. We will say that a feasible outcome is **Pareto optimal** if there is no other feasible outcome that both players weakly prefer and one player strictly prefers. If an effort level e is part of a Pareto optimal outcome, we will say that it is a **first-best** effort level, and we will denote it by e^{FB}.

Lemma 6. The first-best effort level satisfies

$$e^{FB} \in \underset{e \in \mathcal{E}}{\operatorname{argmax}} \, E_y \left[py \middle| \, e \right] - c\left(e\right).$$

Proof of Lemma 6. In any Pareto-optimal outcome, payments to the agent are deterministic. Since the Agent is risk averse, given an outcome involving stochastic payments to the Agent, there is another outcome in which the Agent chooses the same effort level and receives the certainty equivalent wage instead. This outcome yields the same utility for the Agent, and since the Agent is risk averse, the certainty equivalent payment is smaller in expectation, so the Principal is strictly better off. Next, given constant deterministic wages, any Pareto-optimal outcome must solve

$$\max_{w \in \mathbb{R}, e \in \mathcal{E}} E_y \left[py \middle| e \right] - w$$

subject to

$$u \left(w - c \left(e \right) \right) \geq \bar{u},$$

for some $\bar{u}$. In any solution to this problem, the constraint must bind, since u is increasing. Moreover, since u is increasing, it is invertible, so we can write

$$w = u^{-1} \left(\bar{u} \right) + c \left(e \right),$$

and therefore the first-best effort level must solve the problem specified in the Lemma.■

This Lemma shows that the first-best effort level maximizes expected revenues net of effort costs. If effort is fully contractible, so that the Principal could offer any contract w that depended nontrivially on e, then the first-best effort would be implemented in equilibrium. In particular, the Principal could offer a contract that pays the Agent $u^{-1} \left(\bar{u} \right) + c \left(e^{FB} \right)$ if he choose e^{FB}, and pays him a large negative amount if he chooses any $e \neq e^{FB}$. That the first-best effort level can be implemented in equilibrium if effort is contractible is an illustration of a version of the *Coase Theorem*: if the contracting space is sufficiently rich, equilibrium outcomes will be Pareto optimal.

If effort is noncontractible, and $e^{FB} > 0$, then equilibrium will not involve

Pareto optimal outcomes. For an outcome to be Pareto optimal, it has to involve a deterministic wage payment to the Agent. But if the Agent's wage is independent of output, then it must also be independent of his effort level. He will therefore receive no benefit from choosing a costly effort level, and so he will choose $e = 0 < e^{FB}$. The question to which we will now turn is: what effort will be implemented in equilibrium when effort is noncontractible?

Analysis Since the Agent's effort choice affects the Principal's payoffs, the Principal would ideally like to directly choose the Agent's effort. But, she has only indirect control: she can offer different contracts, and different contracts may get the Agent to optimally choose different effort levels. We can think of the Principal's problem as choosing an effort level e as well as a contract for which e is *incentive compatible* for the Agent to choose and for which it is *individually rational* for the Agent to accept. As a loose analogy, we can connect the Principal's problem to the social planner's problem from general equilibrium theory. We can think of e as analogous to an allocation the Principal would like to induce, and the choice of a contract as analogous to setting "prices" so as to decentralize e as an equilibrium allocation.

Formally, the Principal offers a contract $w \in \mathcal{W}$ and "proposes" an effort level e in order to solve

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e)$$

subject to two constraints. The first constraint is that the agent actually prefers to choose effort level e rather than any other effort level $\hat{e}$. This is the **incentive-compatibility constraint**:

$$e \in \underset{\hat{e} \in \mathcal{E}}{\operatorname{argmax}} \int_{y \in Y} u\left(w\left(y\right) - c\left(\hat{e}\right)\right) dF\left(y \middle| \hat{e}\right).$$

The second constraint ensures that, given that the agent knows he will choose e if he accepts the contract, he prefers to accept the contract rather than to reject it and receive his outside utility $\bar{u}$. This is the **individual-rationality constraint** or **participation constraint**:

$$\int_{y \in Y} u\left(w\left(y\right) - c\left(e\right)\right) dF\left(y \middle| e\right) \geq \bar{u}.$$

At this level of generality, the model is not very tractable. We will need to impose more structure on it in order to highlight some its key trade-offs and properties.

CARA-Normal Case with Affine Contracts In order to highlight one of the key trade-offs that arise in this class of models, we will make a number of strong simplifying assumptions.

Assumption A1 (CARA). The Agent has CARA preferences over wealth and effort costs, which are quadratic:

$$u\left(w\left(y\right) - c\left(e\right)\right) = -\exp\left\{-r\left(w\left(y\right) - \frac{c}{2}e^2\right)\right\},$$

and his outside option yields utility $-\exp\{-r\bar{u}\}$.

Assumption A2 (Normal Output). Effort shifts the mean of a normally distributed random variable. That is, $y \sim N(e, \sigma^2)$.

Assumption A3 (Affine Contracts). $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}, w(y) = s + by\}$. That is, the contract space permits only affine contracts.

Assumption A4 (Continuous Effort). Effort is continuous and satisfies $\mathcal{E} = \mathbb{R}_+$.

In principle, we should not impose exogenous restrictions on the *functional form* of $w(y)$. There is an important class of applications, however, that restrict attention to affine contracts, $w(y) = s+by$, and a lot of the basic intuition that people have for the comparative statics of optimal contracts come from imposing this restriction.

In many environments, an optimal contract does not exist if the contracting space is sufficiently rich, and situations in which the agent chooses the first-best level of effort, and the principal receives all the surplus can be arbitrarily approximated with a sequence of sufficiently perverse contracts (Mirrlees, 1974; Moroni and Swinkels, 2014). In contrast, the optimal affine contract often results in an effort choice that is lower than the first-best effort level, and the principal receives a lower payoff.

There are then at least three ways to view the exercise of solving for the optimal affine contract.

1. From an applied perspective, many pay-for-performance contracts in

the world are affine in the relevant performance measure—franchisees pay a franchise fee and receive a constant fraction of the revenues their store generates, windshield installers receive a base wage and a constant piece rate, fruit pickers are paid per kilogram of fruit they pick. And so given that many practitioners seem to restrict attention to this class of contracts, why not just make sure they are doing what they do optimally? Put differently, we can brush aside global optimality on purely pragmatic grounds.

2. Many pay-for-performance contracts in the world are affine in the relevant performance measure. Our models are either too rich or not rich enough in a certain sense and therefore generate optimal contracts that are inconsistent with those we see in the world. Maybe the aspects that, in the world, lead practitioners to use affine contracts are orthogonal to the considerations we are focusing on, so that by restricting attention to the optimal affine contract, we can still say something about how real-world contracts ought to vary with changes in the underlying environment. This view presumes a more positive (as opposed to normative) role for the modeler and hopes that the theoretical analogue of the omitted variables bias is not too severe.

3. Who cares about second-best when first-best can be attained? If our models are pushing us toward complicated, non-linear contracts, then maybe our models are wrong. Instead, we should focus on writing

> down models that generate affine contracts as the optimal contract, and therefore we should think harder about what gives rise to them. (And indeed, steps have been made in this direction—see Holmström and Milgrom (1987), Diamond (1998) and, more recently, Carroll (2015) and Barron, Georgiadis, and Swinkels (2017)) This perspective will come back later in the course when we discuss the Property Rights Theory of firm boundaries.

Given Assumptions $(A1)-(A3)$, for any contract $w(y)=s+by$, the income stream the agent receives is normally distributed with mean $s+be$ and variance $b^2\sigma^2$. His expected utility over monetary compensation is therefore a moment-generating function for a normally distributed random variable, (recall that if $X\sim N(\mu,\sigma^2)$, then $E[\exp\{tX\}]=\exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}$), so his preferences can be written as

$$E[-\exp\{-r(w(y)-c(e))\}]=-\exp\left\{-r\left(s+be-\frac{r}{2}b^2\sigma^2-\frac{c}{2}e^2\right)\right\}.$$

We can take a monotonic transformation of his utility function ($f(x)=-\frac{1}{r}\log(-x)$) and represent his preferences as:

$$\begin{aligned} U(e,w) &= E[w(y)]-\frac{r}{2}Var(w(y))-\frac{c}{2}e^2 \\ &= s+be-\frac{r}{2}b^2\sigma^2-\frac{c}{2}e^2. \end{aligned}$$

The Principal's program is then

$$\max_{s,b,e} pe - (s + be)$$

subject to incentive-compatibility

$$e \in \underset{\hat{e}}{\operatorname{argmax}}\, b\hat{e} - \frac{c}{2}\hat{e}^2$$

and individual-rationality

$$s + be - \frac{r}{2}b^2\sigma^2 - \frac{c}{2}e^2 \geq \bar{u}.$$

Solving this problem is then relatively straightforward. Given an affine contract $s + be$, the Agent will choose an effort level $e(b)$ that satisfies his first-order conditions

$$e(b) = \frac{b}{c},$$

and the Principal will choose the value s to ensure that the Agent's individual-rationality constraint holds with equality. If it did not hold with equality, the Principal could reduce s, making herself better off without affecting the Agent's incentive-compatibility constraint, while still respecting the Agent's individual-rationality constraint. That is,

$$s + be(b) = \frac{c}{2}e(b)^2 + \frac{r}{2}b^2\sigma^2 + \bar{u}.$$

In other words, the Principal has to ensure that the Agent's total expected monetary compensation, $s + be\left(b\right)$, fully compensates him for his effort costs, the risk costs he has to bear if he accepts this contract, and his opportunity cost. Indirectly, then, the Principal bears these costs when designing an optimal contract.

The Principal's remaining problem is to choose the incentive slope b to solve

$$\max_{b} pe\left(b\right) - \frac{c}{2}e\left(b\right)^2 - \frac{r}{2}b^2\sigma^2 - \bar{u}.$$

This is now an unconstrained problem with proper convexity assumptions, so the Principal's optimal choice of incentive slope solves her first-order condition

$$0 = p\underbrace{e'\left(b^*\right)}_{1/c} - c\underbrace{e^*\left(b^*\right)}_{b^*/c}\underbrace{e'\left(b^*\right)}_{1/c} - rb^*\sigma^2,$$

and therefore the optimal incentive slope satisfies

$$b^* = \frac{p}{1 + rc\sigma^2}.$$

Moreover, given b^* and the individual-rationality constraint, we can back out s^*.

$$s^* = \bar{u} + \frac{1}{2}\left(rc\sigma^2 - 1\right)\frac{\left(b^*\right)^2}{c}.$$

Depending on the parameters, it may be the case that $s^* < 0$. That is, the Agent would have to pay the Principal if he accepts the job and does not produce anything.

Now, how does the effort that is induced in this optimal affine contract compare to the **first-best effort**? Using the result from Lemma 1, we know that first-best effort in this setting solves

$$\max_{e\in\mathbb{R}_+} pe - \frac{c}{2}e^2,$$

and therefore $e^{FB} = p/c$.

Even if effort is noncontractible, the Principal could in principle implement exactly this same level of effort by writing a contract only on output. To do so, she would choose $b = p$, since this would get the Agent to choose $e\left(p\right) = p/c$. Why, in this setting, does the Principal not choose such a contract? Let us go back to the Principal's problem of choosing the incentive slope b.

$$\max_b pe\left(b\right) - \frac{c}{2}e\left(b\right)^2 - \frac{r}{2}b^2\sigma^2 - \bar{u}$$

Often, when an economic model can be solved in closed form, we jump right to the solution. Only when a model cannot be solved in closed form do we typically stop to think carefully about what economic properties its solution must possess. I want to spend a couple minutes *partially* characterizing this model's solution, even though we already completely characterized it above, just to highlight how this kind of reasoning can be helpful in developing intuition that might generalize beyond the present setting. In particular, many fundamental features of models can be seen as a comparison of first-order losses or gains against second-order gains or losses, so it is worth going through this first-order–second-order logic. Suppose the Principal chooses $b = p$, and consider a marginal reduction in b away from this value. The change in the Principal's profits would be

$$\begin{aligned} & \frac{d}{db}\left(pe\left(b\right)-\frac{c}{2}e\left(b\right)^{2}-\frac{r}{2}b^{2}\sigma^{2}\right)\bigg|_{b=p} \\ = \; & \underbrace{\frac{d}{db}\left(pe\left(b\right)-\frac{c}{2}e\left(b\right)^{2}\right)\bigg|_{b=p}}_{=0}-rp\sigma^{2}<0. \end{aligned}$$

This first term is zero, because $b = p$ in fact maximizes $pe\left(b\right)-\frac{c}{2}e\left(b\right)^{2}$, since it induces the first-best level of effort. This is just an application of the envelope theorem you learned in Ec 2010a. The second term in this expression is strictly negative. This implies that, relative to the contract that induces first-best effort, a reduction in the slope of the incentive contract yields a first-order gain to the Principal resulting from a decrease in the risk costs the Agent bears, while it yields a second-order loss in terms of profits resulting from moving away from the effort level that maximizes revenues minus effort costs. The optimal contract balances the incentive benefits of higher-powered incentives with these risk costs, and these risk costs are higher if the Agent is more risk averse and if output is noisier.

This trade-off seems first-order in some settings (e.g., insurance contracts in health care markets, some types of sales contracts in industries in which individual sales are infrequent, large, and unpredictable) and for certain types of output. There are many other environments in which contracts provide less-than-first-best incentives, but the first-order reasons for these low-powered contracts seem completely different, and we will turn to these environments next week.

Exercise 18 (Adapted from MWG 14.B.4). Suppose there are three possible effort levels, $\mathcal{E} = \{e_1, e_2, e_3\}$, and two possible output levels, $\mathcal{Y} = \{0, 10\}$, and the output price is $p = 1$. The probability that $y = 10$ conditional on each of the effort levels is given by the probability mass function $f(10|\, e_1) = 2/3$, $f(10|\, e_2) = 1/2$, and $f(10|\, e_3) = 1/3$. The Agent's effort cost function satisfies $c(e_1) = 5/3$, $c(e_2) = 8/5$, and $c(e_3) = 4/3$. Finally, the Agent's utility function is given by $u(w) = \sqrt{w}$, and his outside option yields utility $\bar{u} = 0$.

(a) What is the optimal contract for the Principal when effort is contractible?

(b) Show that if effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$, then there is no contract w for which the Agent will choose e_2. For what levels of $c(e_2)$ would there exist a contract w under which the Agent would choose e_2?

(c) What is the optimal contract when effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$?

(d) Suppose instead that $c(e_1) = \sqrt{8}$, and let $f(10|\, e_1) = x \in (0, 1)$. If effort is noncontractible, and $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}\}$, what is the optimal contract for the Principal as x approaches 1? Is the level of effort implemented higher or lower than when effort is contractible?

Exercise 19. Suppose the Agent can allocate time to two different tasks. Let e_i be the amount of time spent on task $i \in \{1, 2\}$. The Principal cares only about task 1 and obtains payoff $y = e_1 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. The Agent, however, derives a benefit $v(e_2)$ from spending time on task 2. The Agent has CARA preferences with utility function

$$u(w, e_1, e_2) = -\exp\{-r[w - c(e_1 + e_2) + v(e_2)]\},$$

where $c(e_1 + e_2)$ is the cost of time, with $c'(\cdot) > 0$, $c''(\cdot) > 0$, and $c(0) = 0$. Assume also that $v'(\cdot) > 0$, $v''(\cdot) < 0$, and $v(0) = 0$, and that optimization with respect to (e_1, e_2) results in an interior solution. Let $\bar{w}$ denote the wage the Agent receives from his outside option, so $\bar{u} = -\exp\{-r\bar{w}\}$.

(a) What is the first-best outcome in this setting?

(b) Suppose effort e_1 is noncontractible, and the Principal can write a contract that is an affine function of output and can also allow the Agent to engage in task 2 or not. Under these assumptions, what is the contracting space?

(c) Suppose the Principal must pay the Agent $s = 1$ if $y = 0$. Will the Principal allow the Agent to engage in task 2? Compare this to your answer in part (a). What if $s < 1$ is set exogenously? Find the difference in the Principal's utility under the two policies, as a function of s.

Exercise 20. This exercise goes through a two-period version of Holmström and Milgrom's (1987) linear contracts argument. In each of two periods, $t \in \{1, 2\}$, the Agent chooses whether to "work" or to "shirk": $e_t \in \{0, 1\}$ at cost ce_t with $c > 0$. Output is binary, so that $y_t \in \{0, 1\}$, and the price of output is normalized to 1. Effort increases the probability that $y_t = 1$:

$$1 > \Pr\left[y_t = 1 |\, e_t = 1\right] = p_H > p_L = \Pr\left[y_t = 1 |\, e_t = 0\right] > 0.$$

The Agent's Bernoulli utility function is

$$u\left(w, e_1, e_2\right) = -\exp\left\{-r\left(w - ce_1 - ce_2\right)\right\},$$

and his outside option yields utility $-\exp\{-r \cdot 0\}$. The Agent can observe the realization of y_1 before choosing y_2.

The Principal's payoff is $y_1 + y_2 - w$, and the payment w can depend on each period's output and is paid at the end of period 2 (i.e., after both realizations of output). Assume it is optimal to induce the Agent to work hard in both periods. Show that a least-cost (optimal) contract that implements $e_1 = e_2 = 1$ has the form

$$w\left(y_1, y_2\right) = s + b\left(y_1 + y_2\right).$$

Guide:

(a) Define w_{y_1,y_2} to be the wage conditional on y_1 in period 1 and y_2 in period 2. Then, using the (IC) constraints for period 2, show that

$$e^{rc}\left[1 + p_H\left\{\frac{\exp\{-rw_{0,1}\}}{\exp\{-rw_{0,0}\}} - 1\right\}\right] = 1 + p_L\left\{\frac{\exp\{-rw_{0,1}\}}{\exp\{-rw_{0,0}\}} - 1\right\}$$

and

$$e^{rc}\left[1+p_H\left\{\frac{\exp\{-rw_{1,1}\}}{\exp\{-rw_{1,0}\}}-1\right\}\right]=1+p_L\left\{\frac{\exp\{-rw_{1,1}\}}{\exp\{-rw_{1,0}\}}-1\right\}.$$

This implies there exists

$$b=w_{0,1}-w_{0,0}=w_{1,1}-w_{1,0}.$$

Why did CARA utility matter for this argument?

(b) Now, using the (IC) constraint for period 1, show that we have

$$e^{rc}\left[1+p_H\left\{\frac{u_1}{u_0}-1\right\}\right]=1+p_L\left\{\frac{u_1}{u_0}-1\right\},$$

where u_i is the expected utility conditional on success in the first period $(i=1)$ or failure $(i=0)$.

(c) Note that

$$\exp\{r(c-w_{y,1})\}=\exp\{-rb\}\exp\{r(c-w_{y,0})\}$$

for each $y\in\{0,1\}$. Now show that

$$\frac{u_1}{u_0}=\exp\{-r(w_{1,0}-w_{0,0})\}=\exp\{-r(w_{1,1}-w_{0,1})\}.$$

Therefore, we must have $b=w_{0,1}-w_{0,0}=w_{1,0}-w_{0,0}=w_{1,1}-w_{0,1}$.

The First-Order Approach

Last time, we imposed a lot of structure on the Principal-Agent problem and solved for optimal affine contracts. One of the problems we identified with that approach was that there was not a particularly compelling reason for restricting attention to affine contracts. Moreover, in that particular setting, if we allowed the contracts to take more general functional forms, there in fact was no optimal contract.

Today, we will return to a slightly modified version of the more general setup of the problem and consider an alternative approach to characterizing optimal contracts without imposing any assumptions on the functional forms they might take. One change we will be making is that the Agent's preferences are now given by

$$U(w,e)=\int_{y\in\mathcal{Y}}\left[u(w(y))-c(e)\right]dF(y|e)=E_y\left[u(w)|e\right]-c(e),$$

where u is strictly increasing and strictly concave, and the utility the Agent receives from money is additively separable from his effort costs.

Recall from last time that the Principal's problem is to choose an output-contingent contract $w\in\mathcal{W}\subset\{w:\mathcal{Y}\to\mathbb{R}\}$ and to "propose" an effort level e to solve:

$$\max_{w\in\mathcal{W},e\in\mathcal{E}}\int_{y\in\mathcal{Y}}(py-w(y))\,dF(y|e)$$

subject to an incentive-compatibility constraint

$$e\in\underset{\hat{e}\in\mathcal{E}}{\operatorname{argmax}}\int_{y\in\mathcal{Y}}u(w(y))\,dF(y|\hat{e})-c(\hat{e})$$

and an individual-rationality constraint

$$\int_{y\in\mathcal{Y}}u(w(y))\,dF(y|e)-c(e)\geq\bar{u}.$$

One of the problems with solving this problem at this level of generality is that the incentive-compatibility constraint is quite a complicated set of conditions. The contract has to ensure that, of all the effort levels the Agent could potentially choose, he prefers to choose e. In other words, the contract has to deter the Agent from choosing any other effort level $\hat{e}$: for all $\hat{e} \in \mathcal{E}$, we must have

$$\int_{y \in \mathcal{Y}} [u(w(y)) - c(e)] \, dF(y|\, e) \geq \int_{y \in \mathcal{Y}} [u(w(y)) - c(\hat{e})] \, dF(y|\, \hat{e}).$$

When effort is continuous, the incentive-compatibility constraint is actually a continuum of constraints of this form. It seems like it should be the case that if we impose more structure on the problem, we can safely ignore most of these constraints. This turns out to be true. If we impose some relatively stringent but somewhat sensible assumptions on the problem, then if it is the case that the Agent does not want to deviate *locally* to another $\hat{e}$, then he also does not want to deviate to an $\hat{e}$ that is farther away. When local constraints are sufficient, we will in fact be able to replace the Agent's incentive-compatibility constraint with the first-order condition to his problem.

Throughout, we will be focusing on models that satisfy the following assumptions.

Assumption A1 (Continuous Effort and Continuous Output). Effort is continuous and satisfies $\mathcal{E} = \mathbb{R}_+$. Output is continuous, with $\mathcal{Y} = \mathbb{R}$, and for each $e \in \mathcal{E}$, $F(\cdot|\, e)$ has support $\left[\underline{y}, \bar{y}\right]$ and has density $f(\cdot|\, e)$, where $f(\cdot|\, e)$ is differentiable in e.

Assumption A2 (First-Order Stochastic Dominance—FOSD). The output distribution function satisfies $F_e(y|\, e) \leq 0$ for all $e \in \mathcal{E}$ and all y with strict inequality for some y for each e.

Assumption $(A2)$ roughly says that higher effort levels make lower output realizations less likely and higher output realizations more likely. This assumption provides sufficient conditions under which higher effort increases total expected surplus, ignoring effort costs.

We will first explore the implications of being able to replace the incentive-compatibility constraint with the Agent's first-order condition, and then we will provide some sufficient conditions under which doing so is without loss of generality. Under Assumption $(A1)$, if we replace the Agent's incentive-compatibility constraint with his first-order condition, the Principal's problem becomes:

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{\underline{y}}^{\bar{y}} (py - w(y))\, f(y|\, e)\, dy$$

subject to the local incentive-compatibility constraint

$$c'(e) = \int_{\underline{y}}^{\bar{y}} u(w(y))\, f_e(y|\, e)\, dy$$

and the individual-rationality constraint

$$\int_{\underline{y}}^{\bar{y}} u\left(w\left(y\right)\right) f\left(\left.y\right| e\right) dy - c\left(e\right) \geq \bar{u}.$$

This problem is referred to as the **first-order approach** to characterizing second-best incentive contracts. It is now just a constrained-optimization problem with an equality constraint and an inequality constraint. We can therefore write the Lagrangian for this problem as

$$\begin{aligned} \mathcal{L} &= \int_{\underline{y}}^{\bar{y}} \left(py - w\left(y\right)\right) f\left(\left.y\right| e\right) dy + \lambda \left(\int_{\underline{y}}^{\bar{y}} u\left(w\left(y\right)\right) f\left(\left.y\right| e\right) dy - c\left(e\right) - \bar{u} \right) \\ & +\mu \left(\int_{\underline{y}}^{\bar{y}} u\left(w\left(y\right)\right) f_e\left(\left.y\right| e\right) dy - c'\left(e\right) \right), \end{aligned}$$

where λ is the Lagrange multiplier on the individual-rationality constraint, and μ is the Lagrange multiplier on the local incentive-compatibility constraint. We can derive the conditions for the optimal contract $w^*\left(y\right)$ inducing optimal effort e^* by taking first-order conditions, point-by-point, with respect to $w\left(y\right)$. These conditions are:

$$\frac{1}{u'\left(w^*\left(y\right)\right)} = \lambda + \mu \frac{f_e\left(\left.y\right| e^*\right)}{f\left(\left.y\right| e^*\right)}.$$

Contracts satisfying these conditions are referred to as Holmström-Mirrlees contracts (or (λ, μ) contracts as one of my colleagues calls them). There are several points to notice here. First, the left-hand side is increasing in $w\left(y\right)$, since u is concave. Second, if $\mu = 0$, then this condition would correspond to the conditions for an optimal risk-sharing rule between the Principal and the Agent. Under a Pareto-optimal risk allocation, the **Borch Rule** states that the ratio of the Principal's marginal utility to the Agent's marginal utility is equalized across states. In this case, the Principal's marginal utility is one. Any optimal-risk sharing rule will equalize the Agent's marginal utility of income across states and therefore give the Agent a constant wage.

Third, Holmström (1979) shows that under Assumption $(A2)$, $\mu > 0$, so that the right-hand side of this equation is increasing in $f_e(y|e^*)/f(y|e^*)$. You might remember from econometrics that this ratio is called the **score**—it tells us how an increase in e changes the log likelihood of e given output realization y. To prevent the Agent from choosing effort level e instead of e^*, the contract has to pay the Agent more for outputs that are more likely under e^* than under e. Since by assumption, we are looking at only local incentive constraints, the contract will pay the Agent more for outputs that are more likely under e^* than under effort levels arbitrarily close to e^*.

Together, these observations imply that the optimal contract $w^*(y)$ is increasing in the score. Just because an optimal contract is increasing in the score does not mean that it is increasing in output. The following assumption guarantees that the score is increasing in y, and therefore optimal contracts are increasing in output.

Assumption A3 (Monotone Likelihood Ratio Property—MLRP). Given any two effort levels $e, e' \in \mathcal{E}$ with $e > e'$, the ratio $f(y|e)/f(y|e')$ is increasing in y.

MLRP guarantees, roughly speaking, that higher levels of output are more indicative of higher effort levels.[1] Under Assumption $(A1)$, MLRP is equivalent to the condition that $f_e\left(y|\, e\right)/f\left(y|\, e\right)$ is increasing in y. We can therefore interpret the optimality condition as telling us that the optimal contract is increasing in output precisely when higher output levels are more indicative of higher effort levels. Put differently, the optimal contract "wants" to reward *informative* output, not necessarily *high* output.

The two statistical properties, FOSD and MLRP, that we have assumed come up a lot in different settings, and it is easy to lose track of what they each imply. To recap, the FOSD property tells us that higher effort makes higher output more likely, and it guarantees that there is always a benefit of higher effort levels, gross of effort costs. The MLRP property tells us that higher output is more indicative of higher effort, and it guarantees that optimal contracts are increasing in output. These two properties are related: MLRP implies FOSD, but not the reverse.

Informativeness Principle

Before we provide conditions under which the first-order approach is valid, we will go over what I view as the most important result to come out of this

[1] The property can also be interpreted in terms of statistical hypothesis testing. Suppose the null hypothesis is that the Agent chose effort level e', and the alternative hypothesis is that the Agent chose effort level $e > e'$. If, given output realization y, a likelihood ratio test would reject the null hypothesis of lower effort, the same test would also reject the null hypothesis for any higher output realization.

model. Suppose there is another contractible performance measure $m \in \mathcal{M}$, where y and m have joint density function $f(y, m|\, e)$, and the contracting space is $\mathcal{W} = \{w : \mathcal{Y} \times \mathcal{M} \to \mathbb{R}\}$. Under what conditions will an optimal contract $w(y, m)$ depend nontrivially on m? The answer is: whenever m provides additional information about e. To make this argument precise, we will introduce the following definition.

Definition 14. Given two random variables Y and M, Y is **sufficient for** (Y, M) **with respect to** $e \in \mathcal{E}$ if and only if the joint density function $f(y, m|\, e)$ is multiplicatively separable in m and e:

$$f(y, m|\, e) = g(m|\, e)\, h(y, m).$$

We will say that M is **informative about** $e \in \mathcal{E}$ if Y is not sufficient for (Y, M) with respect to $e \in \mathcal{E}$.

We argued above that optimal contracts pay the Agent more for outputs that are more indicative of high effort. This same argument also extends to other performance measures, as long as they are informative about effort. This result is known as the *informativeness principle* and was first established by Holmström (1979) and Shavell (1979).

Theorem 15 (Informativeness Principle). Assume the first-order approach is valid. Let $w(y)$ be the optimal contract when m is noncontractible. If m is contractible, there exist a contract $w(y, m)$ that Pareto dominates $w(y)$ if and only if m is informative about $e \in \mathcal{E}$.

Proof of Theorem 15. In both cases, the optimal contract gives the Agent $\bar{u}$, so we just need to show that the Principal can be made strictly better off if m is contractible.

If the first-order approach is valid, the optimality conditions for the Principal's problem when both y and m are contractible are given by

$$\frac{1}{u'\left(w^*\left(y,m\right)\right)} = \lambda + \mu \frac{f_e\left(y,m\right|e^*)}{f\left(y,m\right|e^*)}.$$

The optimal contract $w^*\left(y,m\right)$ is independent of m if and only if y is sufficient for $\left(y,m\right)$ with respect to e^*.

This result seems like it should be obvious: optimal contracts clearly should make use of all available information. But it is not ex ante obvious this would be the case. In particular, one could easily have imagined that optimal contracts should only depend on performance measures that are "sufficiently" informative about effort—after all, basing a contract on another performance measure could introduce additional noise as well. Or one could have imagined that optimal contracts should only depend on performance measures that are directly affected by the Agent's effort choice. The informativeness principle says that optimal contracts should depend on every performance measure that is even slightly informative.

This result has both positive and negative implications. On the positive and practical side, it says that optimal contracts should make use of benchmarks: a fund manager should be evaluated for her performance relative to a market index, CEOs should be rewarded for firm performance relative to other firms in their industry, and employees should be evaluated relative to their peers. On the negative side, the result shows that optimal contracts are highly sensitive to the fine details of the environment. This implication is, in a real sense, a weakness of the theory: it is the reason why the theory often predicts contracts that bear little resemblance to what we actually see in practice.

The informativeness principle was derived under the assumption that the first-order approach was valid. When the first-order approach is not valid, the informativeness principle does not necessarily hold. The reason for this is that when the first-order approach does not hold, there may be multiple binding incentive-compatibility constraints at the optimum, and just because an informative performance measure helps relax one of those constraints, if it does not help relax the other binding constraints, it need not strictly increase the firm's profits. Chaigneau, Edmans, and Gottlieb (forthcoming) generalizes the informativeness principle to settings in which the first-order approach is not valid.

Validity of the First-Order Approach

Finally, we will briefly talk about some sufficient conditions ensuring the first-order approach is valid. Assumption $(A4)$, along with the following assumption, are sufficient.

Assumption A4 (Convexity of the Distribution Function Condition—

CDFC). $F(\cdot|\, e)$ is twice differentiable, and $F_{ee}(\cdot|\, e) \geq 0$ for all e.

CDFC is a strong assumption. There is a fairly standard class of distributions that are often used in contract theory that satisfy it, but it is not satisfied by other well-known families of distributions. Let $F_H(y)$ and $F_L(y)$ be two distribution functions that have density functions $f_H(y)$ and $f_L(y)$ for which $f_H(y)/f_L(y)$ is increasing in y, and suppose

$$F(y|\, e) = eF_H(y) + (1-e)F_L(y).$$

Then $F(y|\, e)$ satisfies both MLRP and CDFC. In other words, MLRP and CDFC are satisfied if output is drawn from a mixture of a "high" and a "low" distribution, and higher effort increases the probability that output is drawn from the high distribution.

Theorem 16. Suppose $(A1)-(A4)$ are satisfied. If the local incentive-compatibility constraint is satisfied, the incentive-compatibility constraint is satisfied.

Proof sketch of Theorem 16. The high-level idea of the proof is to show that MLRP and CDFC imply that the Agent's effort-choice problem is globally concave for any contract the Principal offers him. Using integration by parts, we can rewrite the Agent's expected utility as follows.

$$\begin{aligned}\int_{\underline{y}}^{\bar{y}} u(w(y)) f(y|e)\, dy - c(e) &= u(w(y)) F(y|e)|_{\underline{y}}^{\bar{y}} \\ &\quad - \int_{\underline{y}}^{\bar{y}} u'(w(y)) \frac{dw(y)}{dy} F(y|e)\, dy - c(e) \\ &= u(w(\bar{y})) - \int_{\underline{y}}^{\bar{y}} u'(w(y)) \frac{dw(y)}{dy} F(y|e)\, dy - c(e).\end{aligned}$$

Now, suppose $w(y)$ is increasing and differentiable. Differentiating the expression above with respect to e twice yields

$$-\int_{\underline{y}}^{\bar{y}} u'(w(y)) \frac{dw(y)}{dy} F_{ee}(y|e)\, dy - c''(e) < 0$$

for every $e \in \mathcal{E}$, since $F_{ee} > 0$. Thus, the Agent's second-order condition is globally satisfied, so if the local incentive constraint is satisfied, the incentive constraint is satisfied.■

I labeled this proof as a sketch, because while it follows Mirrlees's (1976) argument, the full proof (due to Rogerson (1985)) requires showing that $w(y)$ is in fact increasing and differentiable when MLRP is satisfied. We cannot use our argument above for why MLRP implies increasing contracts, because that argument presumed the first-order approach was valid, which is exactly what we are trying to prove here. The MLRP and CDFC conditions are known as the Mirrlees-Rogerson conditions.

There are other sufficient conditions for the first-order approach to be valid that do not require such strong distributional assumptions (see, for example, Jewitt (1988)). And there are other approaches to solving the moral hazard problem that do not rely on the first-order approach. These include Grossman and Hart (1983), which decomposes the Principal's problem into two steps: the first step solves for the cost-minimizing contract that implements a given effort level, and the second step solves for the optimal effort level. We will take this approach when we think about optimal contracts under limited liability in the next section.

4.2 Limited Liability and Incentive Rents

We saw in the previous model that the optimal contract sometimes involved up-front payments from the Agent to the Principal. To the extent that the Agent is unable to afford such payments (or legal restrictions such as minimum wage laws prohibit such payments), the Principal will not be able to extract all the surplus that the Agent creates. Further, in order to extract surplus from the Agent, the Principal may have to put in place contracts that reduce the total surplus created. In equilibrium, the Principal may therefore offer a contract that induces effort below the first-best.

The Model Again, there is a risk-neutral Principal (P). There is also a **risk-neutral** Agent (A). The Agent chooses an effort level $e \in \mathcal{E} \subset \mathbb{R}_+$ at a cost of $c(e)$, where $c : \mathbb{R}_+ \to \mathbb{R}_+$, with $c'', c' > 0$, and this effort level affects the distribution over outputs $y \in \mathcal{Y}$, with y distributed according to CDF $F(\cdot|\, e)$. These outputs can be sold on the product market for price p. The Principal can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{Y} \to \mathbb{R}, w(y) \geq \underline{w} \text{ for all } y\}$ that determines a transfer $w(y)$ that she is compelled to pay the Agent if output y is realized. The Agent has an outside option that provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the outside option is not exercised, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned} \Pi(w, e) &= \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e) = E_y[py - w|\, e] \\ U(w, e) &= \int_{y \in \mathcal{Y}} (w(y) - c(e))\, dF(y|\, e) = E_y[w - c(e)|\, e]. \end{aligned}$$

There are two differences between this model and the previous model. The first difference is that the Agent is risk-neutral (so that absent any other changes, the equilibrium contract would induce first-best effort). The second difference is that the wage payment from the Principal to the Agent has to exceed, for each realization of output, a value $\underline{w}$. Depending on the setting, this constraint is described as a liquidity constraint or a limited-liability constraint. In repeated settings, it is more naturally thought of as the latter—due to legal restrictions, the Agent cannot be legally compelled to make a transfer (larger than $-\underline{w}$) to the Principal. In static settings, either interpretation may be sensible depending on the particular application—if the Agent is a fruit picker, for instance, he may not have much liquid wealth that he can use to pay the Principal.

Timing The timing of the game is exactly the same as before.

1. P offers A a contract $w(y)$, which is commonly observed.

2. A accepts the contract $(d = 1)$ or rejects it $(d = 0)$ and receives $\bar{u}$, and the game ends. This decision is commonly observed.

3. If A accepts the contract, A chooses effort level e and incurs cost $c(e)$. e is only observed by A.

4. Output y is drawn from distribution with cdf $F(\cdot|\, e)$. y is commonly observed.

5. P pays A an amount $w(y)$. This payment is commonly observed.

Equilibrium The solution concept is the same as before. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0, 1\}$, and an effort choice $e^* : \mathcal{W} \times \{0, 1\} \to \mathbb{R}_+$ such that given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract induces effort e^*.

The Program The Principal offers a contract $w \in \mathcal{W}$ and proposes an effort level e in order to solve

$$\max_{w \in \mathcal{W}, e \in \mathcal{E}} \int_{y \in \mathcal{Y}} (py - w(y))\, dF(y|\, e)$$

subject to three constraints: the incentive-compatibility constraint

$$e \in \underset{\hat{e} \in \mathcal{E}}{\operatorname{argmax}} \int_{y \in \mathcal{Y}} \left(w\left(y\right) - c\left(\hat{e}\right)\right) dF\left(y\right| \hat{e}\right),$$

the individual-rationality constraint

$$\int_{y \in \mathcal{Y}} \left(w\left(y\right) - c\left(e\right)\right) dF\left(y\right| e\right) \geq \bar{u},$$

and the limited-liability constraint

$$w\left(y\right) \geq \underline{w} \text{ for all } y \in \mathcal{Y}.$$

Binary-Output Case We will impose much more structure on the problem to illustrate the main trade-off in this class of models. Innes (1990) and Jewitt, Kadan, and Swinkels (2008) explore a much more general analysis.

Assumption A1 (Binary Output). Output is $y \in \{0, 1\}$, and given effort e, its distribution satisfies $\Pr\left[y = 1\right| e\right] = e$.

Assumption A2 (Well-behaved Cost). The Agent's costs have a nonnegative third derivative: $c''' \geq 0$, and they satisfy conditions that ensure an interior solution: $c'\left(0\right) = 0$ and $c'\left(1\right) = +\infty$. Or for comparison across models in this module, $c\left(e\right) = \frac{c}{2}e^2$, where $p \leq c$ to ensure that $e^{FB} < 1$.

Finally, we can restrict attention to affine, nondecreasing contracts

$$\begin{aligned} \mathcal{W} &= \{w(y) = (1-y)\,w_0 + yw_1, w_1 \geq w_0 \geq 0\} \\ &= \{w(y) = s + by, s \geq \underline{w}, b \geq 0\}\,. \end{aligned}$$

When output is binary, this restriction to affine contracts is without loss of generality. Also, the restriction to nondecreasing contracts is not restrictive (i.e., any optimal contract of a relaxed problem in which we do not impose that contracts are nondecreasing will also be the solution to the full problem). This result is something that needs to be shown and is not in general true, but in this case, it is straightforward.

As Grossman and Hart (1983) highlight, in Principal–Agent models, it is often useful to break the problem down into two steps. The first step takes a target effort level, e, as given and solves for the set of cost-minimizing contracts implementing effort level e. Any cost-minimizing contract implementing effort level e results in an expected cost of $C(e)$ to the principal. The second step takes the function $C(\cdot)$ as given and solves for the optimal effort choice.

In general, the cost-minimization problem tends to be a well-behaved convex-optimization problem, since (even if the agent is risk-averse) the objective function is weakly concave, and the constraint set is a convex set (since given an effort level e, the individual-rationality constraint and the limited-liability constraint define convex sets, and each incentive constraint ruling out effort level $\hat{e} \neq e$ also defines a convex set, and the intersection of convex sets is itself a convex set). The resulting cost function $C(\cdot)$ need not have nice properties, however, so the second step of the optimization problem is only well-behaved under restrictive assumptions. In the present case, Assumptions $(A1)$ and $(A2)$ ensure that the second step of the optimization problem is well-behaved.

Cost-Minimization Problem Given an effort level e, the cost-minimization problem is given by

$$C(e, \bar{u}, \underline{w}) = \min_{s,b} s + be$$

subject to the Agent's incentive-compatibility constraint

$$e \in \underset{\hat{e}}{\operatorname{argmax}} \{s + b\hat{e} - c(\hat{e})\},$$

his individual-rationality constraint

$$s + be - c(e) \geq \bar{u},$$

and the limited-liability constraint

$$s \geq \underline{w}.$$

I will denote a **cost-minimizing contract implementing effort level** e by (s_e^*, b_e^*).

The first step in solving this problem is to notice that the Agent's incentive-compatibility constraint implies that any cost-minimizing contract implementing effort level e must have $b_e^* = c'(e)$.

If there were no limited-liability constraint, the Principal would choose s_e^* to extract the Agent's surplus. That is, given $b = b_e^*$, s would solve

$$s + b_e^* e = \bar{u} + c(e) .$$

That is, s would ensure that the Agent's expected compensation exactly equals his expected effort costs plus his opportunity cost. The resulting s, however, may not satisfy the limited-liability constraint. The question then is: given $\bar{u}$ and $\underline{w}$, for what effort levels e is the Principal able to extract all the agent's surplus (i.e., for what effort levels does the limited-liability constraint not bind at the cost-minimizing contract?), and for what effort levels is she unable to do so? Figure 15 below shows cost-minimizing contracts for effort levels e_1 and e_2. Any contract can be represented as a line in this figure, where the line represents the expected pay the Agent will receive given an effort level e. The cost-minimizing contract for effort level e_1 is tangent to the $\bar{u} + c(e)$ curve at e_1 and its intercept is $s_{e_1}^*$. Similarly for e_2. Both $s_{e_1}^*$ and $s_{e_2}^*$ are greater than $\underline{w}$, which implies that for such effort levels, the limited-liability constraint is not binding.

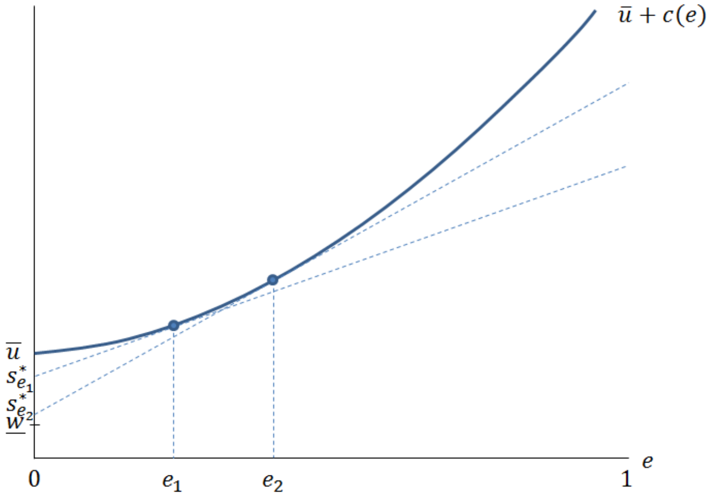


Figure 15: Cost-minimizing contracts

For effort sufficiently high, the limited-liability constraint will be binding in a cost-minimizing contract, and it will be binding for all higher effort levels. Define the threshold $\bar{e}\left(\bar{u},\underline{w}\right)$ to be the effort level such that for all $e \geq \bar{e}\left(\bar{u},\underline{w}\right)$, $s_e^* = \underline{w}$. Figure 16 illustrates that $\bar{e}\left(\bar{u},\underline{w}\right)$ is the effort level at which the contract tangent to the $\bar{u}+c\left(e\right)$ curve at $\bar{e}\left(\bar{u},\underline{w}\right)$ intersects the vertical axis at exactly $\underline{w}$. That is, $\bar{e}\left(\bar{u},\underline{w}\right)$ solves

$$c'\left(\bar{e}\left(\bar{u},\underline{w}\right)\right) = \frac{\bar{u}+c\left(\bar{e}\left(\bar{u},\underline{w}\right)\right)-\underline{w}}{\bar{e}\left(\bar{u},\underline{w}\right)}.$$

Figure 2 also illustrates that for all effort levels $e > \bar{e}\left(\bar{u},\underline{w}\right)$, the cost-minimizing contract involves giving the Agent strictly positive surplus. That is, the cost to the Principal of getting the agent to choose effort $e > \bar{e}\left(\bar{u}, \underline{w}\right)$ is equal to the Agent's opportunity costs $\bar{u}$ plus his effort costs $c\left(e\right)$ plus **incentive costs** $IC\left(e, \bar{u}, \underline{w}\right)$.

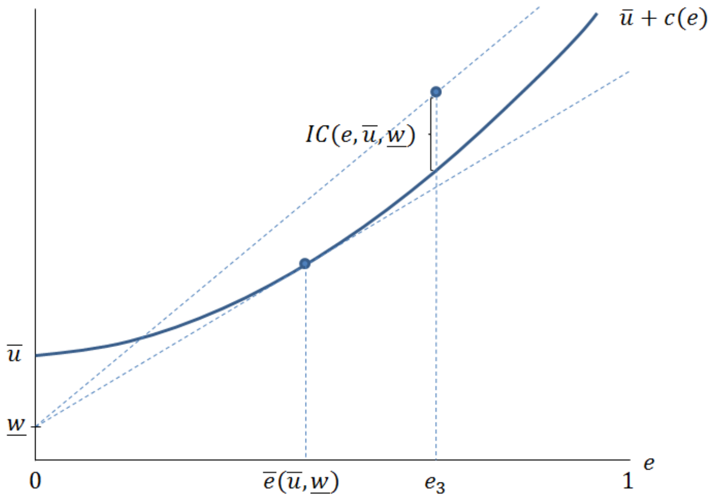


Figure 16: Incentive Costs for High Effort Levels

The incentive costs $IC\left(e, \bar{u}, \underline{w}\right)$ are equal to the Agent's expected compensation given effort choice e and cost-minimizing contract $\left(s_e^*, b_e^*\right)$ minus his costs:

$$\begin{aligned} IC\left(e, \bar{u}, \underline{w}\right) &= \begin{cases} 0 & e \leq \bar{e}\left(\bar{u}, \underline{w}\right) \\ \underline{w} + c'\left(e\right)e - c\left(e\right) - \bar{u} & e \geq \bar{e}\left(\bar{u}, \underline{w}\right) \end{cases} \\ &= \max\left\{0, \underline{w} + c'\left(e\right)e - c\left(e\right) - \bar{u}\right\} \end{aligned}$$

where I used the fact that for $e \geq \bar{e}(\bar{u}, \underline{w})$, $s_e^* = \underline{w}$ and $b_e^* = c'(e)$. This incentive-cost function $IC(\cdot, \bar{u}, \underline{w})$ is the key object that captures the main contracting friction in this model. I will sometimes refer to $IC(e, \bar{u}, \underline{w})$ as the **incentive rents** required to get the Agent to choose effort level e. Putting these results together, we see that

$$C(e, \bar{u}, \underline{w}) = \bar{u} + c(e) + IC(e, \bar{u}, \underline{w}).$$

That is, the Principal's total costs of implementing effort level e are the sum of the Agent's costs plus the incentive rents required to get the Agent to choose effort level e.

Since $IC(e, \bar{u}, \underline{w})$ is the main object of interest in this model, I will describe some of its properties. First, it is continuous in e (including, in particular, at $e = \bar{e}(\bar{u}, \underline{w})$). Next, $\bar{e}(\bar{u}, \underline{w})$ and $IC(e, \bar{u}, \underline{w})$ depend on $(\bar{u}, \underline{w})$ only inasmuch as $(\bar{u}, \underline{w})$ determines $\bar{u} - \underline{w}$, so I will abuse notation and write these expressions as $\bar{e}(\bar{u} - \underline{w})$ and $IC(e, \bar{u} - \underline{w})$. Also, given that $c'' > 0$, IC is increasing in e (since $\underline{w} + c'(e)e - c(e) - \underline{u}$ is strictly increasing in e, and IC is just the max of this expression and zero). Further, given that $c''' \geq 0$, IC is convex in e. For $e \geq \bar{e}(\bar{u} - \underline{w})$, this property follows, because

$$\frac{\partial^2}{\partial e^2} IC = c''(e) + c'''(e)\, e \geq 0.$$

And again, since IC is the max of two convex functions, it is also a convex function. Finally, since $IC(\cdot, \bar{u} - \underline{w})$ is flat when $e \leq \bar{e}(\bar{u} - \underline{w})$ and it is strictly increasing (with slope independent of $\bar{u}-\underline{w}$) when $e \geq \bar{e}(\bar{u}-\underline{w})$, the slope of IC with respect to e is (weakly) decreasing in $\bar{u}-\underline{w}$, since $\bar{e}(\bar{u}-\underline{w})$ is increasing in $\bar{u}-\underline{w}$. That is, $IC(e,\bar{u}-\underline{w})$ satisfies decreasing differences in $(e,\bar{u}-\underline{w})$.

Motivation-Rent Extraction Trade-off The second step of the optimization problem takes as given the function

$$C(e,\bar{u}-\underline{w}) = \bar{u} + c(e) + IC(e,\bar{u}-\underline{w})$$

and solves the Principal's problem for the optimal effort level:

$$\begin{aligned} & \max_e pe - C(e,\bar{u}-\underline{w}) \\ = \; & \max_e pe - \bar{u} - c(e) - IC(e,\bar{u}-\underline{w}). \end{aligned}$$

Note that total surplus is given by $pe - \bar{u} - c(e)$, which is therefore maximized at $e = e^{FB}$ (which, if $c(e) = ce^2/2$, then $e^{FB} = p/c$). Figure 17 below depicts the Principal's expected benefit line pe, and her expected costs of implementing effort e at minimum cost, $C(e,\bar{u}-\underline{w})$. The first-best effort level, e^{FB} maximizes the difference between pe and $\bar{u}+c(e)$, while the equilibrium effort level e^* maximizes the difference between pe and $C(e,\bar{u}-\underline{w})$.

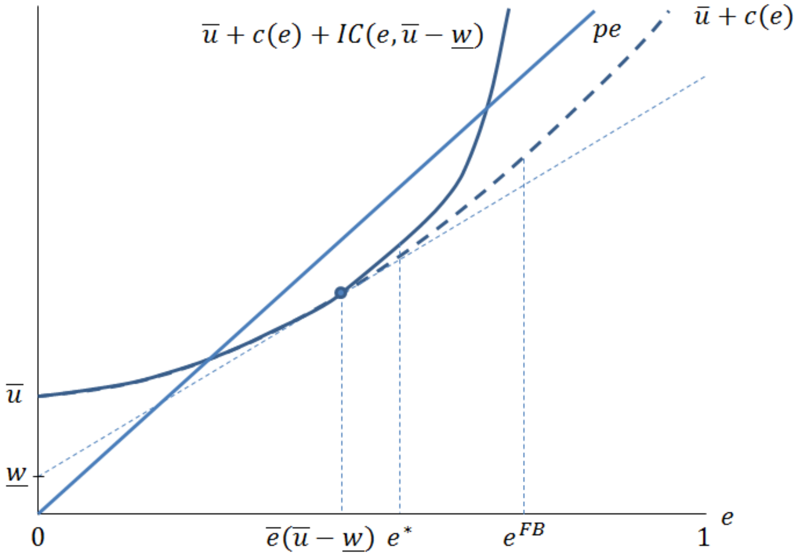


Figure 17: Optimal Effort Choice

If $c(e) = ce^2/2$, we can solve explicitly for $\bar{e}(\bar{u} - \underline{w})$ and for $IC(e, \bar{u} - \underline{w})$ when $e > \bar{e}(\bar{u} - \underline{w})$. In particular,

$$\bar{e}(\bar{u} - \underline{w}) = \left(\frac{2(\bar{u} - \underline{w})}{c}\right)^{1/2}$$

and when $e > \bar{e}(\bar{u} - \underline{w})$,

$$IC(e, \bar{u} - \underline{w}) = \underline{w} + \frac{1}{2}ce^2 - \bar{u}.$$

If $\underline{w} < 0$ and p is sufficiently small, we can have $e^* = e^{FB}$ (i.e., these are the conditions required to ensure that the limited-liability constraint is not binding for the cost-minimizing contract implementing $e = e^{FB}$). If p is sufficiently large relative to $\bar{u} - \underline{w}$, we will have $e^* = \frac{1}{2}\frac{p}{c} = \frac{1}{2}e^{FB}$. For p somewhere in between, we will have $e^* = \bar{e}\left(\bar{u} - \underline{w}\right) < e^{FB}$. In particular, $C\left(e, \bar{u} - \underline{w}\right)$ is kinked at this point.

As in the risk–incentives model, we can illustrate through a partial characterization why (and when) effort is less-than first-best. Since we know that e^{FB} maximizes $pe - \bar{u} - c\left(e\right)$, we therefore have that

$$\frac{d}{de}\left[pe - \bar{u} - c\left(e\right) - IC\left(e, \bar{u} - \underline{w}\right)\right]_{e=e^{FB}} = -\frac{\partial}{\partial e} IC\left(e^{FB}, \bar{u} - \underline{w}\right) \leq 0,$$

with strict inequality if the limited-liability constraint binds at the cost-minimizing contract implementing e^{FB}. This means that, even though e^{FB} maximizes total surplus, if the Principal has to provide the agent with rents at the margin, she may choose to implement a lower effort level. Reducing the effort level away from e^{FB} leads to second-order losses in terms of total surplus, but it leads to first-order gains in profits for the Principal. In this model, there is a tension between total-surplus creation and rent extraction, which yields less-than-first-best effort in equilibrium.

In my view, liquidity constraints are extremely important and are probably one of the main reasons for why many jobs do not involve first-best incentives. The logic that first-best efforts can be implemented if the firm transfers the entire profit stream to each of its members in exchange for a large up-front payment seems simultaneously compelling, trivial, and obviously impracticable. In for-profit firms, in order to make it worthwhile to transfer a large enough share of the profit stream to an individual worker to significantly affect his incentives, the firm would require a large up-front transfer that most workers cannot afford to pay. It is therefore not surprising that we do not see most workers' compensation tied directly to the firm's overall profits in a meaningful way. One implication of this logic is that firms have to find alternative instruments to use as performance measures, which we will turn to next. In principle, models in which firms do not motivate their workers by writing contracts directly on profits should include assumptions under which the firm optimally chooses not to write contracts directly on profits, but they almost never do.

Exercise 21. This exercise goes through a version of Diamond's (1998) and Barron, Georgiadis, and Swinkels's (2018) argument for why linear contracts are optimal when the Agent is able to "take on risk." Suppose the Principal and the Agent are both risk neutral, and let $\mathcal{Y} = [0, \bar{y}]$ and $\mathcal{E} = \mathbb{R}_+$. There is a limited-liability constraint, and the contracting space is $\mathcal{W} = \{w : \mathcal{Y} \to \mathbb{R}_+\}$. After the Agent chooses an effort level e, he can then choose any distribution function $F(y)$ over output that satisfies $e = \int_0^{\bar{y}} y dF(y)$. In other words, his effort level determines his *average* output, but he can then add mean-preserving noise to his output. Given a contract w, effort e, and distribution F, the Agent's expected utility is

$$\int_0^{\bar{y}} w(y) \, dF(y) - c(e),$$

where c is strictly increasing and strictly convex. The Principal's expected profits are $\int_0^{\bar{y}} (y - w(y)) \, dF(y)$. The Agent's outside option gives both parties a payoff of zero.

(a) Show that a linear contract of the form $w(y) = by$ maximizes the Principal's expected profits. To do so, you will want to argue that given any contract $w(y)$ that implements effort level e, there is a linear contract that also implements effort level e but at a weakly lower cost to the Principal. [Hint: instead of thinking about all the possible distribution functions the Agent can choose among, it may be useful to just look at distributions that put weight on two levels of output, $0 \leq y_L < y_H \leq \bar{y}$ satisfying $e = (1-q) y_L + q y_H$.]

(b) Are there other contracts that maximize the Principal's expected profits? If so, how are they related to the optimal linear contract? If not, provide an intuition for why linear contracts are uniquely optimal.

4.3 Misaligned Performance Measures

In the previous two models, the Principal cared about output, and output, though a noisy measure of effort, was perfectly measurable. This assumption seems sensible when we think about overall firm profits (ignoring basically everything that accountants think about every day), but as we alluded to in the previous discussion, overall firm profits are too blunt of an instrument to use to motivate individual workers within the firm if they are liquidity-constrained. As a result, firms often try to motivate workers using more specific performance measures, but while these performance measures are informative about what actions workers are taking, they may be less useful as a description of how the workers' actions affect the objectives the firm cares about. And paying workers for what is measured may not get them to take actions that the firm cares about. This observation underpins the title of the famous 1975 paper by Steve Kerr called "On the Folly of Rewarding A, While Hoping for B."

As an example, think of a retail firm that hires an employee both to make sales and to provide customer service. It can be difficult to measure the quality of customer service that a particular employee provides, but it is easy to measure that employee's sales. Writing a contract that provides the employee with high-powered incentives directly on sales will get him to put a lot of effort into sales and very little effort into customer service. And in fact, he might only be able to put a lot of effort into sales by intentionally neglecting customer service. If the firm cares equally about both dimensions, it might be optimal not to offer high-powered incentives to begin with. This is what Holmström and Milgrom (1991) refers to as the "multitask problem." We will look at a model that captures some of this intuition, although not as directly as Holmström and Milgrom's model. The model we will look at builds upon Baker (1992, 2002) and Feltham and Xie (1994).

Description Again, there is a risk-neutral Principal (P) and a risk-neutral Agent (A). The Agent chooses an effort vector $e = (e_1, e_2) \in \mathcal{E} \subset \mathbb{R}^2_+$ at a cost of $\frac{c}{2}\left(e_1^2 + e_2^2\right)$. This effort vector affects the distribution of output $y \in \mathcal{Y} = \{0, 1\}$ and a performance measure $m \in \mathcal{M} = \{0, 1\}$ as follows:

$$\begin{aligned} \Pr\left[y = 1 \middle| e\right] &= f_1 e_1 + f_2 e_2 \\ \Pr\left[m = 1 \middle| e\right] &= g_1 e_1 + g_2 e_2, \end{aligned}$$

where it may be the case that $f = (f_1, f_2) \neq (g_1, g_2) = g$. Assume that $f_1^2 + f_2^2 = g_1^2 + g_2^2 = 1$ (i.e., the norms of the f and g vectors are unity). The output can be sold on the product market for price p. Output is noncontractible, but the performance measure is contractible. The Principal can write a contract $w \in \mathcal{W} \subset \{w : \mathcal{M} \to \mathbb{R}\}$ that determines a transfer $w(m)$ that she is compelled to pay the Agent if performance measure m is realized. Since the performance measure is binary, contracts take the form $w = s + bm$. The Agent has an outside option that provides utility $\bar{u}$ to the Agent and $\bar{\pi}$ to the Principal. If the outside option is not exercised, the Principal's and Agent's preferences are, respectively,

$$\begin{aligned} \Pi(w, e) &= f_1 e_1 + f_2 e_2 - s - b(g_1 e_1 + g_2 e_2) \\ U(w, e) &= s + b(g_1 e_1 + g_2 e_2) - \frac{c}{2}\left(e_1^2 + e_2^2\right). \end{aligned}$$

Timing The timing of the game is exactly the same as before.

1. P offers A a contract w, which is commonly observed.

2. A accepts the contract $(d = 1)$ or rejects it $(d = 0)$ and receives $\bar{u}$ and the game ends. This decision is commonly observed.

3. If A accepts the contract, A chooses effort vector e. e is only observed by A.

4. Performance measure m and output y are drawn from the distributions described above. m is commonly observed.

5. P pays A an amount $w(m)$. This payment is commonly observed.

Equilibrium The solution concept is the same as before. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0,1\}$, and an effort choice $e^* : \mathcal{W} \times \{0,1\} \to \mathbb{R}_+^2$ such that given the contract w^*, the Agent optimally chooses d^* and e^*, and given d^* and e^*, the Principal optimally offers contract w^*. We will say that the optimal contract induces effort e^*.

The Program The principal offers a contract w and proposes an effort level e to solve

$$\max_{s,b,e} p\left(f_1 e_1 + f_2 e_2\right) - \left(s + b\left(g_1 e_1 + g_2 e_2\right)\right)$$

subject to the incentive-compatibility constraint

$$e \in \underset{\hat{e} \in \mathbb{R}_+^2}{\operatorname{argmax}}\, s + b\left(g_1 \hat{e}_1 + g_2 \hat{e}_2\right) - \frac{c}{2}\left(\hat{e}_1^2 + \hat{e}_2^2\right)$$

and the individual-rationality constraint

$$s + b\left(g_1 e_1 + g_2 e_2\right) - \frac{c}{2}\left(e_1^2 + e_2^2\right) \geq \bar{u}.$$

Equilibrium Contracts and Effort Given a contract $s + bm$, the Agent will choose

$$e_1^*(b) = \frac{b}{c}g_1;\ e_2^*(b) = \frac{b}{c}g_2.$$

The Principal will choose s so that the individual-rationality constraint holds with equality

$$s + b\left(g_1 e_1^*(b) + g_2 e_2^*(b)\right) = \bar{u} + \frac{c}{2}\left(e_1^*(b)^2 + e_2^*(b)^2\right).$$

Since contracts send the Agent off in the "wrong direction" relative to what maximizes total surplus, providing the Agent with higher-powered incentives by increasing b sends the agent farther off in the wrong direction. This is costly for the Principal because in order to get the Agent to accept the contract, she has to compensate him for his effort costs, even if they are in the wrong direction.

The Principal's unconstrained problem is therefore

$$\max_b p\left(f_1 e_1^*(b) + f_2 e_2^*(b)\right) - \frac{c}{2}\left(e_1^*(b)^2 + e_2^*(b)^2\right) - \bar{u}.$$

Taking first-order conditions,

$$pf_1 \underbrace{\frac{\partial e_1^*}{\partial b}}_{g_1/c} + pf_2 \underbrace{\frac{\partial e_2^*}{\partial b}}_{g_2/c} = \underbrace{ce_1^*(b^*)}_{b^* g_1/c} \underbrace{\frac{\partial e_1^*}{\partial b}}_{g_1/c} + \underbrace{ce_2^*(b^*)}_{b^* g_2/c} \underbrace{\frac{\partial e_2^*}{\partial b}}_{g_2/c},$$

or

$$b^* = p\frac{f_1 g_1 + f_2 g_2}{g_1^2 + g_2^2} = p\frac{f \cdot g}{g \cdot g} = p\frac{||f||}{||g||}\cos\theta = p\cos\theta,$$

where $\cos\theta$ is the angle between the vectors f and g. That is, the optimal incentive slope depends on the relative magnitudes of the f and g vectors (which in this model were assumed to be the same, but in a richer model this need not be the case) as well as how well-aligned they are. If m is a perfect measure of what the firm cares about, then g is a linear transformation of f and therefore the angle between f and g would be zero, so that $\cos\theta = 1$. If m is completely uninformative about what the firm cares about, then f and g are orthogonal, and therefore $\cos\theta = 0$. As a result, this model is often referred to as the **"cosine of theta model**." (Gibbons, 2010)

It can be useful to view this problem geometrically. Since formal contracts allow for unrestricted lump-sum transfers between the Principal and the Agent, the Principal would optimally like efforts to be chosen in such a way that they maximize total surplus:

$$\max_e p\left(f_1 e_1 + f_2 e_2\right) - \frac{c}{2}\left(e_1^2 + e_2^2\right),$$

which has the same solution as

$$\max_e -\left(e_1 - \frac{p}{c}f_1\right)^2 - \left(e_2 - \frac{p}{c}f_2\right)^2.$$

That is, the Principal would like to choose an effort vector that is collinear with the vector f:

$$\left(e_1^{FB}, e_2^{FB}\right) = \frac{p}{c} \cdot \left(f_1, f_2\right).$$

This effort vector would coincide with the first-best effort vector, since it maximizes total surplus, and the players have quasilinear preferences.

Since contracts can only depend on m and not directly on y, the Principal has only limited control over the actions that the Agent chooses. That is, given a contract specifying incentive slope b, the Agent chooses $e_1^*(b) = \frac{b}{c} g_1$ and $e_2^*(b) = \frac{b}{c} g_2$. Therefore, the Principal can only indirectly "choose" an effort vector that is collinear with the vector g:

$$\left(e_1^*(b), e_2^*(b)\right) = \frac{b}{c} \cdot \left(g_1, g_2\right).$$

The question is then: which such vector maximizes total surplus, which the Principal will extract with an ex ante lump-sum transfer? That is, which point along the $k \cdot (g_1, g_2)$ ray minimizes the mean-squared error distance to $\frac{p}{c} \cdot (f_1, f_2)$?

The following figure illustrates the first-best effort vector e^{FB} and the equilibrium effort vector e^*. The concentric rings around e^{FB} are the Principal's iso-profit curves. The rings that are closer to e^{FB} represent higher profit levels. The optimal contract induces effort vector e^*, which also coincides with the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, g_2)$.

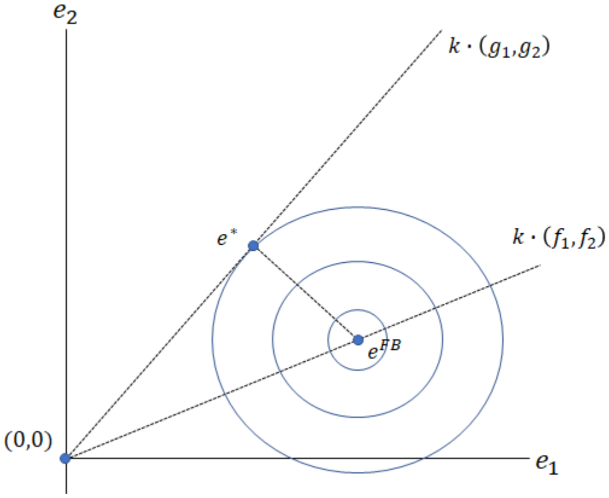


Figure 18: Optimal Effort Vector

This is a more explicit “incomplete contracts” model of motivation. That is, we are explicitly restricting the set of contracts that the Principal can offer the Agent in a way that directly determines a subset of the effort space that the Principal can induce the Agent to choose among. And it is founded not on the idea that certain measures (in particular, y) are unobservable, but rather that they simply cannot be contracted upon.

One observation that is immediate is that it may sometimes be optimal to offer incentive contracts that provide no incentives for the Agent to choose positive effort levels (i.e., $b^* = 0$). This was essentially never the case in the model in which the Agent chose only a one-dimensional effort level, yet we often see that many employees are on contracts that look like they offer no performance-based payments. As this model highlights, this may be optimal precisely when the set of available performance measures are quite bad. As an example, suppose

$$\Pr\left[y = 1 \middle| e\right] = \alpha + f_1 e_1 + f_2 e_2,$$

where $\alpha > 0$ and $f_2 < 0$, so that higher choices of e_2 reduce the probability of high output. And suppose the performance measure is again satisfies

$$\Pr\left[m = 1 \middle| e\right] = g_1 e_1 + g_2 e_2,$$

with $g_1, g_2 > 0$.

We can think of $y = 1$ as representing whether a particular customer buys something that he does not later return, which depends on how well he was treated when he went to the store. We can think of $m = 1$ as representing whether the Agent made a sale but not whether the item was later returned. In order to increase the probability of making a sale, the Agent can exert "earnest" sales effort e_1 and "shady" sales effort e_2. Both are good for sales, but the latter increases the probability the item is returned. If the vectors f and g are sufficiently poorly aligned (i.e., if it is really easy to make sales by being shady), it may be better for the firm to offer a contract with $b^* = 0$, as the following figure illustrates.

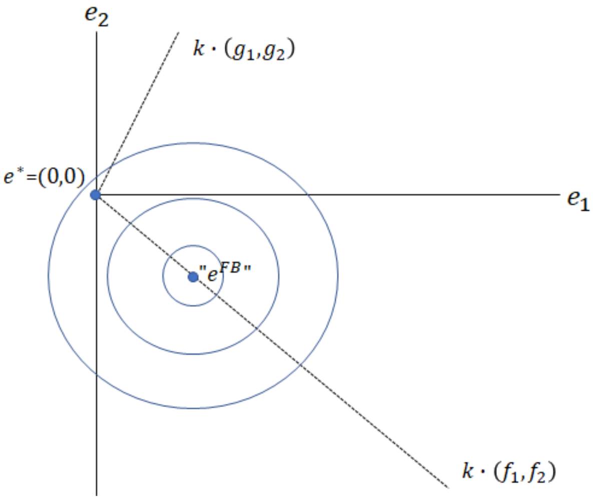


Figure 19: Sometimes Zero Effort is Optimal

This example illustrates that paying the Agent for sales can be a bad idea when what the Principal wants is *sales that are not returned.* The Kerr (1975) article is filled with many colorful examples of this problem. One such example concerns the incentives offered to the managers of orphanages. Their budgets and prestige were determined largely by the number of children they enrolled and not by whether they managed to place their children with suitable families. The claim made in the article is that the managers often denied adoption applications for inappropriate reasons: they were being rewarded for large orphanages, while the state hoped for good placements.

Limits on Activities

Firms have many instruments to help address the problems that arise in multitasking situations. We will describe two of them here in a small extension to the model. Suppose now that the Principal can put some restrictions on the types of actions the Agent is able to undertake. In particular, in addition to writing a contract on the performance measure m, she can write a contract on the dummy variables $1_{e_1>0}$ and $1_{e_2>0}$. In other words, while she cannot directly contract upon, say, e_2, she can write a contract that heavily penalizes any positive level of it. The first question we will ask here is: when does the Principal want to exclude the Agent from engaging in task 2?

We can answer this question using the graphical intuition we just developed above. The following figure illustrates this intuition. If the Principal does not exclude task 2, then she can induce the Agent to choose any effort vector of the form $k \cdot (g_1, g_2)$. If she does exclude task 2, then she can induce the Agent to choose any effort vector of the form $k \cdot (g_1, 0)$. In the former case, the equilibrium effort vector will be e^*, which corresponds to the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, g_2)$. In the latter case, the equilibrium effort will be e^{**}, which corresponds to the orthogonal projection of e^{FB} onto the ray $k \cdot (g_1, 0)$.

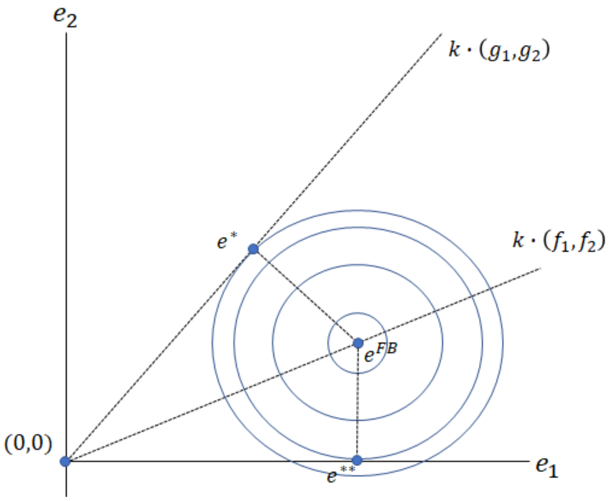


Figure 20: Excluding Task 2

This figure shows that for the particular vectors f and g it illustrates, it will be optimal for the Principal to exclude e_2: e^{**} lies on a higher iso-profit curve than e^* does. This will in fact be the case whenever the angle between vector f and g is larger than the angle between f and $(g_1, 0)$—if by excluding task 2, the performance measure m acts as if it is more closely aligned with f, then task 2 should be excluded.

Job Design

Finally, we will briefly touch upon what is referred to as job design. Suppose f and g are such that it is not optimal to exclude either task on its own. The firm may nevertheless want to hire *two* Agents who each specialize in a single task. For the first Agent, the Principal could exclude task 2, and for the second Agent, the Principal could exclude task 1. The Principal could then offer a contract that gets the first Agent to choose $\left(e_1^{FB}, 0\right)$ and the second agent to choose $\left(0, e_2^{FB}\right)$. The following figure illustrates this possibility.

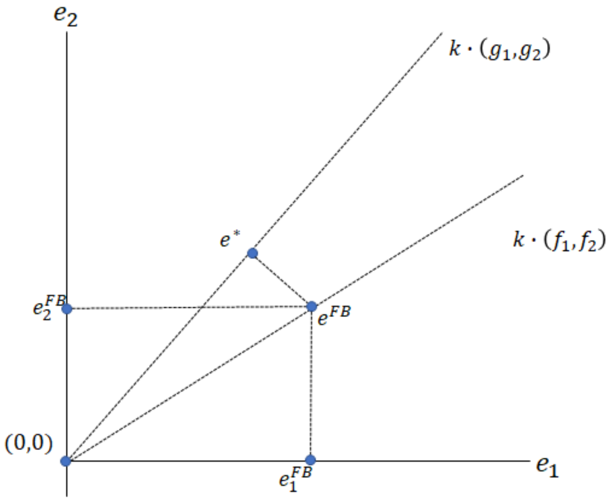


Figure 21: Job Design

When is it optimal for the firm to hire two Agents who each specialize in a single task? It depends on the Agents' opportunity cost. Total surplus under a single Agent under the optimal contract will be

$$pf \cdot e^* - \frac{c}{2} e^* \cdot e^* - \bar{u},$$

and total surplus with two specialized agents under optimal contracts will be

$$pf \cdot e^{FB} - \frac{c}{2} e^{FB} \cdot e^{FB} - 2\bar{u}.$$

Adding an additional Agent in this case is tantamount to adding an additional performance measure, which allows the Principal to choose induce any $e \in \mathbb{R}_{+}^{2}$, including the first-best effort vector. She gains from being able to do this, but to do so, she has to cover the additional Agent's opportunity cost $\bar{u}$.

4.4 Indistinguishable Individual Contributions

So far, we have discussed three contracting frictions that give rise to equilibrium contracts that induce effort that is not first-best. We will now discuss a final contracting friction that arises when multiple individuals contribute to a single project, and while team output is contractible, individual contributions to the team output are not. This indistinguishability gives rise to Holmström's (1982) classic "moral hazard in teams" problem.

The Model There are $I \geq 2$ risk-neutral Agents $i \in \mathcal{I} = \{1, \ldots, I\}$ who each choose efforts $e_i \in \mathcal{E}_i = \mathbb{R}_{+}$ at cost $c_i(e_i)$, which is increasing, convex, differentiable, and satisfies $c_i'(0) = 0$. The vector of efforts $e = (e_1, \ldots, e_I)$ determine team output $y \in \mathcal{Y} = \mathbb{R}_{+}$ according to a function $y(e)$ which is increasing in each e_i, concave in e, differentiable, and satisfies

$\lim_{e_i \to 0} \partial y / \partial e_i = \infty$. Note that output is not stochastic, although the model can be easily extended to allow for stochastic output. Output is contractible, and each Agent i is subject to a contract $w_i \in \mathcal{W} = \{w_i : \mathcal{Y} \to \mathbb{R}\}$. We will say that the vector of contracts $w = (w_1, \ldots, w_I)$ is a **sharing rule** if

$$\sum_{i \in \mathcal{I}} w_i (y) = y$$

for each output level y. Each Agent i's preferences are given by

$$U_i (w, e) = w_i (y (e)) - c_i (e_i) .$$

Each Agent i takes the contracts as given and chooses an effort level. Output is realized and each agent receives payment $w_i (y)$. The solution concept is Nash equilibrium, and we will say that w **induces** e^* if e^* is a Nash equilibrium effort profile given the vector of contracts w.

Sharing Rules and the Impossibility of First-Best Effort Since the Agents have quasilinear preferences, any Pareto-optimal outcome under a sharing rule w will involve an effort level that maximizes total surplus, so that

$$e^{FB} \in \underset{e \in \mathbb{R}_+^I}{\operatorname{argmax}}\, y (e) - \sum_{i \in \mathcal{I}} c_i (e_i) .$$

Under our assumptions, there is a unique first-best effort vector, and it satisfies

$$\frac{\partial y\left(e^{FB}\right)}{\partial e_i} = c_i'\left(e_i^{FB}\right) \text{ for all } i \in \mathcal{I}.$$

First-best effort equates the social marginal benefit of each agent's effort level with its social marginal cost. We will denote the **first-best output level** $y\left(e^{FB}\right)$ by y^{FB}.

We will give an informal argument for why no sharing rule w induces e^{FB}, and then we will make that argument more precise. Suppose w is a sharing rule for which $w_i(y)$ is weakly concave and differentiable in y for all $i \in \mathcal{I}$. For any Nash equilibrium effort vector e^*, it must be the case that

$$w_i'(y) \cdot \frac{\partial y(e^*)}{\partial e_i} = c_i'(e_i^*) \text{ for all } i \in \mathcal{I}.$$

In order for e^* to be equal to e^{FB}, it has to be the case that these equilibrium conditions coincide with the Pareto-optimality conditions. This is only possible if $w_i'(y) = 1$ for all i, but because w is a sharing rule, we must have that

$$\sum_{i \in \mathcal{I}} w_i'(y) = 1 \text{ for all } y.$$

Equilibrium effort e^* therefore cannot be first-best. This argument highlights the idea that getting each Agent to choose first-best effort requires that he be given the entire social marginal benefit of his effort, but it is not possible (at least under a sharing rule) for *all* the Agents simultaneously to receive the entire social marginal benefit of their efforts.

This argument is not a full argument for the impossibility of attaining first-best effort under sharing rules because it does not rule out the possibility of non-differentiable sharing rules inducing first-best effort. It turns out that there is no sharing rule, even a non-differentiable one, that induces first-best effort.

Theorem 16 (Moral Hazard in Teams). If w is a sharing rule, w does not induce e^{FB}.

Proof of Theorem 17. This proof is due to Stole (2001). Take an arbitrary sharing rule w, and suppose e^* is an equilibrium effort profile under w. For any $i, j \in \mathcal{I}$, define $e_j(e_i)$ by the relation $y\left(e_{-j}^*, e_j(e_i)\right) = y\left(e_{-i}^*, e_i\right)$. Since y is continuous and increasing, a unique value of $e_j(e_i)$ exists for e_i sufficiently close to e_i^*. Take such an e_i. For e^* to be a Nash equilibrium, it must be the case that

$$w_j(y(e^*)) - c_j\left(e_j^*\right) \geq w_j\left(y\left(e_{-j}^*, e_j(e_i)\right)\right) - c_j(e_j(e_j)),$$

since this inequality has to hold for all $e_j \neq e_j^*$. Rewriting this inequality, and summing up over $j \in \mathcal{I}$, we have

$$\sum_{j \in \mathcal{I}} \left(w_j(y(e^*)) - w_j\left(y\left(e_{-i}^*, e_i\right)\right)\right) \geq \sum_{j \in \mathcal{I}} \left(c_j\left(e_j^*\right) - c_j(e_j(e_i))\right).$$

Since w is a sharing rule, the left-hand side of this expression is just $y(e^*) -$

$y\left(e_{-i}^{*},e_i\right)$, so this inequality can be written

$$y\left(e^{*}\right)-y\left(e_{-i}^{*},e_i\right)\geq\sum_{j\in\mathcal{I}}c_j\left(e_j^{*}\right)-c_j\left(e_j\left(e_i\right)\right).$$

Since this must hold for all e_i close to e_i^{*}, we can divide by $e_i^{*}-e_i$ and take the limit as $e_i\rightarrow e_i^{*}$ to obtain

$$\frac{\partial y\left(e^{*}\right)}{\partial e_i}\geq\sum_{j\in\mathcal{J}}c_j'\left(e_j^{*}\right)\frac{\partial y\left(e^{*}\right)/\partial e_i}{\partial y\left(e^{*}\right)/\partial e_j}.$$

Now suppose that $e^{*}=e^{FB}$. Then $c_j'\left(e_j^{*}\right)=\partial y\left(e^{*}\right)/\partial e_j$, so this inequality becomes

$$\frac{\partial y\left(e^{*}\right)}{\partial e_i}\geq I\frac{\partial y\left(e^{*}\right)}{\partial e_i},$$

which is a contradiction because y is increasing in e_i.■

Joint Punishments and Budget Breakers Under a sharing rule, first-best effort cannot be implemented because in order to deter an Agent from choosing some $e_i<e_i^{FB}$, it is necessary to punish him. But because contracts can only be written on team output, the only way to deter each agent from choosing $e_i<e_i^{FB}$ is to simultaneously punish *all* the Agents when output is less than $y\left(e^{FB}\right)$. But punishing all the Agents simultaneously requires that they throw output away, which is impossible under a sharing rule. It turns out, though, that if we allow for contracts w that allow for **money burning**,

in the sense that it allows for

$$\sum_{i \in \mathcal{I}} w_i(y) < y$$

for some output levels $y \in \mathcal{Y}$, first-best effort can in fact be implemented, and it can be implemented with a contract that does not actually burn money in equilibrium.

Proposition 4. There exist a vector of contracts w that induces e^{FB} for which $\sum_{i \in \mathcal{I}} w_i\left(y^{FB}\right) = y^{FB}$.

Proof of Proposition 4. For all i, set $w_i(y) = 0$ for all $y \neq y^{FB}$, and let $w_i\left(y^{FB}\right) > c_i\left(e_i^{FB}\right)$ for all i so that $\sum_{i \in \mathcal{I}} w_i\left(y^{FB}\right) = y^{FB}$. Such a vector of contracts is feasible, because $y^{FB} > \sum_{i \in \mathcal{I}} c_i\left(e_i^{FB}\right)$. Finally, under w, e^{FB} is a Nash equilibrium effort profile because if all other Agents choose e_{-i}^{FB}, then if Agent i chooses $e_i \neq e_i^{FB}$, he receives $-c_i(e_i)$, if he chooses $e_i = e_i^{FB}$, he receives $w_i\left(y^{FB}\right) - c_i\left(e_i^{FB}\right) > 0$.■

Proposition 4 shows that in order to induce first-best effort, the Agents have to subject themselves to costly joint punishments in the event that one of them deviates and chooses $e_i \neq e_i^{FB}$. A concern with such contracts is that in the event that the Agents are required by the contract to burn money, they could all be made better off by renegotiating their contract and not burning money. If we insist, therefore, that w is *renegotiation-proof*, then w must be a sharing rule and therefore cannot induce e^{FB}.

This is no longer the case if we introduce an additional party, which we will call a Principal, who does not take any actions that affect output. In particular, if we denote the Principal as Agent 0, then the following sharing rule induces e^{FB}:

$$\begin{aligned} w_i(y) &= y - k \text{ for all } i = 1, \dots, I \\ w_0(y) &= Ik - (I-1)y, \end{aligned}$$

where k satisfies

$$k = \frac{I-1}{I} y^{FB}.$$

This vector of contracts is a sharing rule, since for all $y \in \mathcal{Y}$,

$$\sum_{i=0}^{I} w_i(y) = Iy - (I-1)y = y.$$

This vector of contracts induces e^{FB} because it satisfies $\partial w_i\left(y^{FB}\right)/\partial e_i = 1$ for all $i = 1, \dots, I$, and if we imagine the Principal having an outside option of 0, this choice of k ensures that in equilibrium, she will in fact receive 0. In this case, the Principal's role is to serve as a **budget breaker**. Her presence allows the Agents to "break the margins budget," allowing for $\sum_{i=1}^{I} w_i'(y) = I > 1$, while still allowing for renegotiation-proof contracts.

Under these contracts, the Principal essentially "sells the firm" to *each* agent for an amount k. Then, since each Agent earns the firm's entire output at the margin, each Agent's interests are aligned with society's interest. One limitation of this approach is that while each Agent earns the entire marginal benefit of his efforts, the Principal *loses* $I - 1$ times the marginal benefit of each Agent's efforts. The Principal has strong incentives to collude with one of the Agents—while the players are jointly better off if Agent i chooses e_i^{FB} than any $e_i < e_i^{FB}$, Agent i and the Principal together are jointly better off if Agent i chose $e_i = 0$.

Chapter 5

The Theory of the Firm

The central question in this part of the literature goes back to Ronald Coase (1937): if markets are so great at coordinating productive activity, why is productive activity carried out within firms rather than by self-employed individuals who transact on a spot market? And indeed it is, as Herbert Simon (1991) vividly illustrated:

> A mythical visitor from Mars... approaches Earth from space, equipped with a telescope that reveals social structures. The firms reveal themselves, say, as solid green areas with faint interior contours marking out divisions and departments. Market transactions show as red lines connecting firms, forming a network in the spaces between them. Within firms (and perhaps even between them) the approaching visitor also sees pale blue lines, the lines of authority connecting bosses with various lev-

> els of workers... No matter whether our visitor approached the United States or the Soviet Union, urban China or the European Community, the greater part of the space below it would be within the green areas, for almost all inhabitants would be employees, hence inside the firm boundaries. Organizations would be the dominant feature of the landscape. A message sent back home, describing the scene, would speak of "large green areas interconnected by red lines." It would not likely speak of "a network of red lines connecting green spots." ...When our visitor came to know that the green masses were organizations and the red lines connecting them were market transactions, it might be surprised to hear the structure called a market economy. "Wouldn't 'organizational economy' be the more appropriate term?" it might ask. (pp. 27-28)

It is obviously difficult to put actual numbers on the relative importance of trade within and between firms, since, I would venture to say, most transactions within firms are not recorded. From dropping by a colleague's office to ask for help finding a reference, transferring a shaped piece of glass down the assembly line for installation into a mirror, getting an order of fries from the fry cook to deliver to the customer, most economic transactions are difficult even to define as such, let alone track. But we do have some numbers. The first sentence of Antràs (2003) provides a lower bound: "Roughly one-third of world trade is intrafirm trade."

Of course, it could conceivably be the case that boundaries don't really matter—that the nature of a particular transaction and the overall volume of transactions is the same whether boundaries are in place or not. And indeed, this would exactly be the case if there were no costs of carrying out transactions: Coase's (1960) eponymous theorem suggests, roughly, that in such a situation, outcomes would be the same no matter how transactions were organized. But clearly this is not the case—in 1997, to pick a random year, the volume of corporate mergers and acquisitions was $1.7 trillion dollars (Holmström and Roberts, 1998). It is implausible that this would be the case if boundaries were irrelevant, as even the associated legal fees have to ring up in the billions of dollars.

And so, in a sense, the premise of the Coase Theorem's contrapositive is clearly true. Therefore, there must be transaction costs. And understanding the nature of these transaction costs will hopefully shed some light on the patterns we see. Moreover, as D.H. Robertson vividly illustrated, there are indeed patterns to what we see. Firms are "islands of conscious power in this ocean of unconscious co-operation like lumps of butter coagulating in a pail of buttermilk." So the question becomes: what transaction costs are important, and how are they important? How, in a sense, can they help make sense out of the pattern of butter and buttermilk?

The field was basically dormant for the next forty years until the early 1970s, largely because "transaction costs" came to represent essentially "a name for the residual"—any pattern in the data could trivially be attributed to some story about transaction costs. The empirical content of the theory was therefore essentially zero.

Williamson (1971, 1975, 1979, 1985) put structure on the theory by identifying specific factors that composed these transaction costs. And importantly, the specific factors he identified had implications about economic objects that at least could, in principle, be contained in a data set. Therefore his causal claims could be, and were, tested. (As a conceptual matter, it is important to note that even if Williamson's causal claims were refuted, this would not invalidate the underlying claim that "transaction costs are important," since as discussed earlier, this more general claim is essentially untestable, because it is impossible to measure, or even conceive of, *all* transaction costs associated with *all* different forms of organization.)

The gist of Williamson's Transaction Cost Economics (TCE) theory is that when contracts are incomplete, and parties have disagreements, they may waste resources "haggling" over the appropriate course of action if they transact in a market, whereas if they transact within a firm, these disagreements can be settled by authority or by "fiat." Integration is therefore more appealing than the market when haggling costs are higher, which is the case in situations in which contracts are relatively more incomplete and parties disagree more.

As a classic example (due to Joskow (1985)), think about the relationship between an underground coal mine and a coal fired power plant. It is much more efficient for the power plant to be located close to the coal mine, but the power plant is unlikely to do so absent contractual safeguards. Maybe the parties then end up signing a 20-year contract detailing the type of coal that the mine will send to the power plant and at what price. But after a few years, there may be a regulatory change preventing the use of that particular type of coal. Since such a change is difficult to foresee, the parties may not have specified what to do in this event, and they will have to renegotiate the contract, and these renegotiations may be costly. One way to avoid the problems associated with such renegotiations is vertical integration: the electricity company could buy the coal mine instead of entering into a contract with it. And in the event of a regulatory change, the electricity company just orders the coal mine to produce a different type of coal.

But there was a sense in which TCE theory (and the related work by Klein, Crawford, and Alchian (1978)) was silent on many foundational questions. After all, why does moving the transaction from the market into the firm imply that parties no longer haggle—that is, what is integration? Further, if settling transactions by fiat is more efficient than by haggling, why aren't all transactions carried out within a single firm? Williamson's and others' response was that there are bureaucratic costs ("accounting contrivances," "weakened incentives," and others) associated with putting more transactions within the firm. But surely those costs are also higher when contracts are more incomplete and when there is more disagreement between parties. Put differently, Williamson identified particular costs associated with transacting in the market and other costs associated with transacting within the firm and made assertions about the rates at which these costs vary with the underlying environment. The resulting empirical implications were consistent with evidence, but the theory still lacked convincing foundations, because it treated these latter costs as essentially exogenous and orthogonal.

The Property Rights Theory (PRT), initiated by Grossman and Hart (1986) and expanded upon in Hart and Moore (1990), proposed a theory which (*a*) explicitly answered the question of "what is integration?" and (*b*) treated the costs and benefits of integration symmetrically. Related to the first point is an observation by Alchian and Demsetz that

> It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can "punish" you only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. This is exactly all that any employer can do. He can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products. (1972, p. 777)

What, then, is the difference between me "telling my grocer what to do" and me "telling my employee what to do?" In either case, refusal would potentially cause the relationship to break down. The key difference, according to Grossman and Hart's theory, is in what happens after the relationship breaks down. If I stop buying goods from my grocer, I no longer have access to his store and all its associated benefits. He simply loses access to a particular customer. If I stop employing a worker, on the other hand, the worker loses access to all the assets associated with my firm. I simply lose access to that particular worker.

Grossman and Hart's (1986) key insight is that property rights determine who can do what in the event that a relationship breaks down—property rights determine what they refer to as the residual rights of control. And allocating these property rights to one party or another may change their incentives to take actions that affect the value of this particular relationship. This logic leads to what is often interpreted as Grossman and Hart's main result: property rights (which define whether a particular transaction is carried out "within" a firm or "between" firms) should be allocated to whichever party is responsible for making more important investments in the relationship.

From a theoretical foundations perspective, Grossman and Hart (1986) was a huge step forward—the theory treats the costs of integration and the costs of non-integration symmetrically and systematically analyzes how different factors drive these two costs in a single unified framework. From a conceptual perspective, however, all the action in the theory is related to how organization affects parties' incentives to make relationship-specific investments. As we will see, the theory assumes that conditional on relationship-specific investments, transactions are always carried out efficiently. A manager never wastes time and resources arguing with an employee. An employee never wastes time and resources trying to convince the boss to let him do a different, more desirable task.

Even the Property Rights Theory does not stand on fully firm theoretical grounds, since the theory considers only a limited set of institutions the players can put in place to manage their relationship. That is, PRT focuses only on the allocation of control, ignoring the possibility that individuals may write contracts or put in place other types of mechanisms that could potentially do better. In particular, it rules out revelation mechanisms that, in principle, should induce first-best investment. We will briefly talk about this after we talk about the model.

5.1 Property Rights Theory

Essentially the main result of TCE is the observation that when haggling costs are high under non-integration, then integration is optimal. This result is unsatisfying in at least two senses. First, TCE does not tell us what exactly is the mechanism through which haggling costs are reduced under integration, and second, it does not tell us what the associated costs of integration are, and it therefore does not tell us when we would expect such costs to be high. In principle, in environments in which haggling costs are high under non-integration, then the within-firm equivalent of haggling costs should also be high.

Grossman and Hart (1986) and Hart and Moore (1990) set aside the "make or buy" question and instead begin with the more fundamental question, "What is a firm?" In some sense, nothing short of an answer to *this* question will consistently provide an answer to the questions that TCE leaves unanswered. Framing the question slightly differently, what do I get if I buy a firm from someone else? The answer is typically that I become the owner of the firm's non-human assets.

Why, though, does it matter who owns non-human assets? If contracts are complete, it does not matter. The parties to a transaction will, ex ante, specify a detailed action plan. One such action plan will be optimal. That action plan will be optimal regardless of who owns the assets that support the transaction, and it will be feasible regardless of who owns the assets. If contracts are incomplete, however, not all contingencies will be specified. The key insight of the PRT is that ownership endows the asset's owner with the right to decide what to do with the assets in these contingencies. That is, ownership confers **residual control rights**. When unprogrammed adaptations become necessary, the party with residual control rights has **power** in the relationship and is protected from expropriation by the other party. That is, control over non-human assets leads to control over human assets, since they provide leverage over the person who lacks the assets. Since she cannot be expropriated, she therefore has incentives to make investments that are specific to the relationship.

Firm boundaries are tantamount to asset ownership, so detailing the costs and benefits of different ownership arrangements provides a complete account of the costs and benefits of different firm-boundary arrangements. Asset ownership, and therefore firm boundaries, determine who possesses power in a relationship, and power determines investment incentives. Under integration, I have all the residual control rights over non-human assets and therefore possess strong investment incentives. Non-integration splits apart residual control rights, and therefore provides me with weaker investment incentives and you with stronger investment incentives. If I own an asset, you do not. Power is scarce and therefore should be allocated optimally.

Methodologically, PRT makes significant advances over the preceding theory. PRT's conceptual exercise is to hold technology, preferences, information, and the legal environment constant across prospective governance structures and ask, for a given transaction with given characteristics, whether the transaction is best carried out within a firm or between firms. That is, prior theories associated "make" with some vector $(\alpha_1, \alpha_2, \dots)$ of characteristics and "buy" with some other vector $(\beta_1, \beta_2, \dots)$ of characteristics. "Make" is preferred to "buy" if the vector $(\alpha_1, \alpha_2, \dots)$ is preferred to the vector $(\beta_1, \beta_2, \dots)$. In contrast, PRT focuses on a single aspect: α_1 versus β_1. Further differences may arise between "make" and "buy," but to the extent that they are also choice variables, they will arise optimally rather than by assumption.

The Model There is a risk-neutral upstream manager U, a risk-neutral downstream manager D, and two assets a_1 and a_2. Managers U and D make investments $e_U \in \mathcal{E}_U = \mathbb{R}_+$ and $e_D \in \mathcal{E}_D = \mathbb{R}_+$ at private cost $c_U(e_U)$ and $c_D(e_D)$. These investments determine the value that each manager receives if trade occurs, $V_U(e_U, e_D)$ and $V_D(e_U, e_D)$. There is a state of the world, $s \in \mathcal{S} = \mathcal{S}_C \cup \mathcal{S}_{NC}$, with $\mathcal{S}_C \cap \mathcal{S}_{NC} = \emptyset$ and $\Pr[s \in \mathcal{S}_{NC}] = \mu$. In state s, the identity of the ideal good to be traded is s—if the managers trade good s, they receive $V_U(e_U, e_D)$ and $V_D(e_U, e_D)$. If the managers trade good $s' \neq s$, they both receive $-\infty$. The managers choose an asset allocation, denoted by g, from a set $\mathcal{G} = \{UI, DI, NI, RNI\}$. Under $g = UI$, U owns both assets. Under $g = DI$, D owns both assets. Under $g = NI$, U owns asset a_1 and D owns asset a_2. Under $g = RNI$, D owns asset a_1, and U owns asset a_2. In addition to determining an asset allocation, manager U also offers an incomplete contract $w \in \mathcal{W} = \{w : S_C \to \mathbb{R}\}$ to D. The contract specifies a transfer $w(s)$ to be paid from D to U if they trade good $s \in \mathcal{S}_C$. If the players want to trade a good $s \in \mathcal{S}_{NC}$, they do so in the following way. With probability $\frac{1}{2}$, U makes a take-it-or-leave-it offer $w_U(s)$ to D, specifying trade and a price. With probability $\frac{1}{2}$, D makes a take-it-or-leave-it offer $w_D(s)$ to U specifying trade and a price. If trade does not occur, then manager U receives payoff $v_U(e_U, e_D; g)$ and manager D receives payoff $v_D(e_U, e_D; g)$, which depends on the asset allocation.

Timing There are five periods:

1. U offers D an asset allocation $g \in \mathcal{G}$ and a contract $w \in \mathcal{W}$. Both g and w are commonly observed.

2. U and D simultaneously choose investment levels e_U and e_D at private cost $c\left(e_U\right)$ and $c\left(e_D\right)$. These investment levels are commonly observed by e_U and e_D.

3. The state of the world, $s \in \mathcal{S}$ is realized.

4. If $s \in \mathcal{S}_C$, D buys good s at price specified by w. If $s \in \mathcal{S}_{NC}$, U and D engage in 50-50 take-it-or-leave-it bargaining.

5. Payoffs are realized.

Equilibrium A **subgame-perfect equilibrium** is an asset allocation g^*, a contract w^*, investment strategies $e_U^* : \mathcal{G} \times \mathcal{W} \to \mathbb{R}_+$ and $e_D^* : \mathcal{G} \times \mathcal{W} \to \mathbb{R}_+$, and a pair of offer rules $w_U^* : \mathcal{E}_D \times \mathcal{E}_U \times \mathcal{S}_{NC} \to \mathbb{R}$ and $w_D^* : \mathcal{E}_D \times \mathcal{E}_U \times \mathcal{S}_{NC} \to \mathbb{R}$ such that given $e_U^*\left(g^*, w^*\right)$ and $e_D^*\left(g^*, w^*\right)$, the managers optimally make offers $w_U^*\left(e_U^*, e_D^*\right)$ and $w_D^*\left(e_U^*, e_D^*\right)$ in states $s \in \mathcal{S}_{NC}$; given g^* and w^*, managers optimally choose $e_U^*\left(g^*, w^*\right)$ and $e_D^*\left(g^*, w^*\right)$; and U optimally offers asset allocation g^* and contract w^*.

Assumptions We will assume $c_U\left(e_U\right) = \frac{1}{2}e_U^2$ and $c_D\left(e_D\right) = \frac{1}{2}e_D^2$. We will also assume that $\mu = 1$, so that the probability that an ex ante specifiable good is optimal to trade ex post is zero. Let

$$\begin{aligned} V_U(e_U, e_D) &= f_{UU}e_U + f_{UD}e_D \\ V_D(e_U, e_D) &= f_{DU}e_U + f_{DD}e_D \\ v_U(e_U, e_D; g) &= h^g_{UU}e_U + h^g_{UD}e_D \\ v_D(e_U, e_D; g) &= h^g_{DU}e_U + h^g_{DD}e_D, \end{aligned}$$

and define $F_U = f_{UU} + f_{DU}$ and $F_D = f_{UD} + f_{DD}$. Finally, outside options are more sensitive to one's own investments the more assets one owns:

$$\begin{aligned} h^{UI}_{UU} &\geq h^{NI}_{UU} \geq h^{DI}_{UU}, h^{UI}_{UU} \geq h^{RNI}_{UU} \geq h^{DI}_{UU} \\ h^{DI}_{DD} &\geq h^{NI}_{DD} \geq h^{UI}_{DD}, h^{DI}_{DD} \geq h^{RNI}_{DD} \geq h^{UI}_{DD}. \end{aligned}$$

The Program We solve backwards. For all $s \in \mathcal{S}_{NC}$, with probability $\frac{1}{2}$, U will offer price $w_U(e_U, e_D)$. D will accept this offer as long as $V_D(e_U, e_D) - w_U(e_U, e_D) \geq v_D(e_U, e_D; g)$. U's offer will ensure that this holds with equality (or else U could increase w_U a bit and increase his profits while still having his offer accepted), so that $\pi_U = V_U + V_D - v_D$ and $\pi_D = v_D$.

Similarly, with probability $\frac{1}{2}$, D will offer price $w_D(e_U, e_D)$. U will accept this offer as long as $V_U(e_U, e_D) + w_D(e_U, e_D) \geq v_U(e_U, e_D; g)$. D's offer will ensure that this holds with equality (or else D could decrease w_D a bit and increase her profits while still having her offer accepted), so that $\pi_U = v_U$ and $\pi_D = V_U + V_D - v_U$.

In period 2, manager U will conjecture e_D and solve

$$\max_{\hat{e}_U} \frac{1}{2}\left(V_U\left(\hat{e}_U, e_D\right) + V_D\left(\hat{e}_U, e_D\right) - v_D\left(\hat{e}_U, e_D; g\right)\right) + \frac{1}{2} v_U\left(\hat{e}_U, e_D; g\right) - c\left(\hat{e}_U\right)$$

and manager D will conjecture e_U and solve

$$\max_{\hat{e}_D} \frac{1}{2} v_D\left(e_U, \hat{e}_D; g\right) + \frac{1}{2}\left(V_U\left(e_U, \hat{e}_D\right) + V_D\left(e_U, \hat{e}_D\right) - v_U\left(e_U, \hat{e}_D; g\right)\right) - c\left(\hat{e}_D\right).$$

Given our functional form assumptions, these are well-behaved objective functions, and in each one, there are no interactions between the managers' investment levels, so each manager has a dominant strategy. We can therefore solve for the associated equilibrium investment levels by taking first-order conditions:

$$\begin{aligned} e_U^{*g} &= \frac{1}{2} F_U + \frac{1}{2}\left(h_{UU}^g - h_{DU}^g\right) \\ e_D^{*g} &= \frac{1}{2} F_D + \frac{1}{2}\left(h_{DD}^g - h_{UD}^g\right) \end{aligned}$$

Each manager's incentives to invest are derived from two sources: (1) the marginal impact of investment on total surplus and (2) the marginal impact of investment on the "threat-point differential." The latter point is worth expanding on. If U increases his investment, his outside option goes up by h_{UU}^g, which increases the price that D will have to offer him when she makes her take-it-or-leave-it offer, which increases U's ex-post payoff if $h_{UU}^g > 0$. Further, D's outside option goes up by h_{DU}^g, which increases the price that

U has to offer D when he makes his take-it-or-leave-it-offer, which decreases U's ex-post payoff if $h^g_{DU} > 0$.

Contrasting these equilibrium conditions with the conditions satisfied by first-best effort levels is informative. First-best effort levels satisfy $e_U^{FB} = F_U$ and $e_D^{FB} = F_D$. In contrast, when parties can use renegotiation opportunities to their own advantage, (1) they have weaker incentives to make value-increasing investments that are specific to the relationship, and (2) they may have excessive incentives to make strategic investments in their own outside options or in reducing the outside option of the other party.

Ex ante, players' equilibrium payoffs are:

$$\begin{aligned}
\Pi_U^{*g} &= \frac{1}{2}\left(F_U e_U^{*g} + F_D e_D^{*g}\right) + \frac{1}{2}\left(\left(h^g_{UU} - h^g_{DU}\right) e_U^{*g} + \left(h^g_{UD} - h^g_{DD}\right) e_D^{*g}\right) - \frac{1}{2}\left(e_U^{*g}\right)^2 \\
\Pi_D^{*g} &= \frac{1}{2}\left(F_U e_U^{*g} + F_D e_D^{*g}\right) + \frac{1}{2}\left(\left(h^g_{DU} - h^g_{UU}\right) e_U^{*g} + \left(h^g_{DD} - h^g_{UD}\right) e_D^{*g}\right) - \frac{1}{2}\left(e_D^{*g}\right)^2 .
\end{aligned}$$

If we let $\theta = \left(f_{UU}, f_{UD}, f_{DU}, f_{DD}, \left\{h^g_{UU}, h^g_{UD}, h^g_{DU}, h^g_{DD}\right\}_{g \in G}\right)$ denote the parameters of the model, the Coasian objective for **governance structure** g is:

$$W^g\left(\theta\right) = \Pi_U^{*g} + \Pi_D^{*g} = F_U e_U^{*g} + F_D e_D^{*g} - \frac{1}{2}\left(e_U^{*g}\right)^2 - \frac{1}{2}\left(e_D^{*g}\right)^2 .$$

The **Coasian Problem** that describes the optimal governance structure is then:

$$W^*\left(\theta\right) = \max_{g \in \mathcal{G}} W^g\left(\theta\right).$$

At this level of generality, the model is too rich to provide straightforward insights. In order to make progress, we will introduce the following definitions. If $f_{ij} = h_{ij}^{g} = 0$ for $i \neq j$, we say that investments are **self-investments**. If $f_{ii} = h_{ii}^{g} = 0$, we say that investments are **cross-investments**. When investments are self-investments, the following definitions are useful. Assets A_1 and A_2 are **independent** if $h_{UU}^{UI} = h_{UU}^{NI} = h_{UU}^{RNI}$ and $h_{DD}^{DI} = h_{DD}^{NI} = h_{DD}^{RNI}$ (i.e., if owning the second asset does not increase one's marginal incentives to invest beyond the incentives provided by owning a single asset). Assets A_1 and A_2 are **strictly complementary** if either $h_{UU}^{NI} = h_{UU}^{RNI} = h_{UU}^{DI}$ or $h_{DD}^{NI} = h_{DD}^{RNI} = h_{DD}^{UI}$ (i.e., if for one player, owning one asset provides the same incentives to invest as owning no assets). U's **human capital is essential** if $h_{DD}^{DI} = h_{DD}^{UI}$, and D's human capital is essential if $h_{UU}^{UI} = h_{UU}^{DI}$.

With these definitions in hand, we can get a sense for what features of the model drive the optimal governance-structure choice. (Hart, 1995)

Theorem 18. If A_1 and A_2 are independent, then NI or RNI is optimal. If A_1 and A_2 are strictly complementary, then DI or UI is optimal. If U's human capital is essential, UI is optimal. If D's human capital is essential, DI is optimal. If both U's and D's human capital is essential, all governance structures are equally good.

These results are straightforward to prove. If A_1 and A_2 are independent, then there is no additional benefit of allocating a second asset to a single party. Dividing up the assets therefore strengthens one party's investment incentives without affecting the other's. If A_1 and A_2 are strictly complementary, then relative to integration, dividing up the assets necessarily weakens one party's investment incentives without increasing the other's, so one form of integration clearly dominates. If U's human capital is essential, then D's investment incentives are independent of which assets he owns, so UI is at least weakly optimal.

The more general results of this framework are that (a) allocating an asset to an individual strengthens that party's incentives to invest, since it increases his bargaining position when unprogrammed adaptation is required, (b) allocating an asset to one individual has an opportunity cost, since it means that it cannot be allocated to the other party. Since we have assumed that investment is always socially valuable, this implies that assets should always be allocated to exactly one party (if joint ownership means that both parties have a veto right). Further, allocating an asset to a particular party is more desirable the more important that party's investment is for joint welfare and the more sensitive his/her investment is to asset ownership. Finally, assets should be co-owned when there are complementarities between them.

While the actual results of the PRT model are sensible and intuitive, there are many limitations of the analysis. First, as Holmström (1999) points out, "The problem is that the theory, as presented, really is a theory about asset ownership by individuals rather than by firms, at least if one interprets it literally. Assets are like bargaining chips in an entirely autocratic market... Individual ownership of assets does not offer a theory of organizational identities unless one associates individuals with firms." Holmström concludes that, "... the boundary question is in my view fundamentally about the distribution of activities: What do firms do rather than what do they own? Understanding asset configurations should not become an end in itself, but rather a means toward understanding activity configurations." That is, by taking payoff functions V_U and V_D as exogenous, the theory is abstracting from what Holmström views as the key issue of what a firm really is.

Second, after assets have been allocated and investments made, adaptation is made efficiently. The managers always reach an ex post efficient arrangement in an efficient manner, and all inefficiencies arise ex ante through inadequate incentives to make relationship-specific investments. Williamson (2000) argues that, "The most consequential difference between the TCE and [PRT] setups is that the former holds that maladaptation in the contract execution interval is the principal source of inefficiency, whereas [PRT] vaporize ex post maladaptation by their assumptions of common knowledge and ex post bargaining." That is, Williamson believes that ex post inefficiencies are the primary sources of inefficiencies that have to be managed by adjusting firm boundaries, while the PRT model focuses solely on ex ante inefficiencies. The two approaches are obviously complementary, but there is an entire dimension of the problem that is being left untouched under this approach.

Finally, in the Coasian Problem of the PRT model, the parties are unable to write formal contracts (in the above version of the model, this is true only when $\mu = 1$) and therefore the only instrument they have to motivate relationship-specific investments is the allocation of assets. The implicit assumption underlying the focus on asset ownership is that the characteristics defining what should be traded in which state of the world are difficult to write into a formal contract in a way that a third-party enforcer can unambiguously enforce. State-contingent trade is therefore unverifiable, so contracts written directly or indirectly on relationship-specific investments are infeasible. However, PRT assumes that relationship-specific investments, and therefore the value of different ex post trades, are commonly observable to U and D. Further, U and D can correctly anticipate the payoff consequences of different asset allocations and different levels of investment. Under the assumptions that relationship-specific investments are commonly observable and that players can foresee the payoff consequences of their actions, Maskin and Tirole (1999) shows that the players should always be able to construct a mechanism in which they truthfully reveal the payoffs they would receive to a third-party enforcer. If the parties are able to write a contract on these announcements, then they should indirectly be able to write a contract on ex ante investments. This debate over the "foundations of incomplete contracting" mostly played out over the mid-to-late 1990s, but it has attracted some recent attention.

Exercise 22 (Adapted from Bolton and Dewtripont, Question 42). Consider the following vertical integration problem: there are two risk-neutral managers, each running an asset a_i, where $i = 1, 2$. Both managers make ex ante investments. Only ex post spot contracts regulating trade are feasible. Ex post trade at price P results in the following payoffs: $R(e_D) - P$ for the downstream manager D and $P - C(e_U)$ for the upstream manager U, where the e_i's denote ex ante investment levels. Investing e_U costs the upstream manager e_U, and investing e_D costs the downstream manager e_D.

If the two managers do not trade with each other, their respective payoffs are

$$r\left(e_D, \mathcal{A}_D\right) - P_m \text{ and } P_m - c\left(e_U, \mathcal{A}_U\right),$$

where P_m is a market price, and $\mathcal{A}_i$ denotes the collection of assets owned by manager i. In this problem, $\mathcal{A}_i = \emptyset$ under j-integration, $\mathcal{A}_i = \{a_1, a_2\}$ under i-integration, and $\mathcal{A}_i = \{a_i\}$ under nonintegration.

As in the Grossman-Hart-Moore setting, it is assumed that

$$R\left(e_D\right) - C\left(e_U\right) > r\left(e_D, \mathcal{A}_1\right) - c\left(e_2, \mathcal{A}_2\right)$$

for all $(e_D, e_U) \in [0, \bar{e}]^2$ and all $\mathcal{A}_i$,

$$R'\left(e_D\right) > r'\left(e_D, \{a_1, a_2\}\right) \geq r'\left(e_D, \{a_i\}\right) \geq r'\left(e_D, \emptyset\right) \geq 0,$$

and

$$-C'\left(e_U\right) > -c'\left(e_U, \{a_1, a_2\}\right) \geq -c'\left(e_U, \{a_i\}\right) \geq -c'\left(e_U, \emptyset\right) \geq 0.$$

(a) Characterize the first-best allocation of assets and investment levels.

(b) Assuming that the managers split the ex post gains from trade in half, identify conditions on $r'\left(e_D, \mathcal{A}_i\right)$ and $c'\left(e_D, \mathcal{A}_i\right)$ such that nonintegration is optimal.

Exercise 23. Suppose a downstream buyer D and an upstream seller U meet at date $t = 1$ and trade a widget at date $t = 3$. The value of the widget to the buyer is e_D, and the seller's cost of production is 0. Here, e_D represents an (unverifiable) investment made by the buyer at date $t = 2$. The cost of investment, which is borne entirely by the buyer, is $ce_D^2/2$. No long-term contracts can be written, and there is no discounting.

(a) What is the first-best investment level e_D^{FB}?

(b) Suppose there is a single asset. If the buyer owns it, he has an outside option of λe_D, where $\lambda \in (0, 1)$. If the seller owns it, she has an outside option of v, which is independent of and smaller than e_D. (Imagine that the seller can sell the asset for v in the outside market, and the minimal investment e_D is bigger than v.) Assume that the buyer and seller divide the ex post gains from trade $50 : 50$ (Nash bargaining).

Compute the buyer's investment for the case where the buyer owns the asset and for the case where the seller owns the asset.

(c) Now assume a different bargaining game at date $t = 3$. If both parties have outside options that are valued below $e_D/2$, the parties split the surplus, giving $e_D/2$ to each party. If one of the parties has an outside option that gives $r > e_D/2$, then the party gets r and the other party gets the remainder $e_D - r$. Supposing that $\lambda > 1/2$, compute the buyer's investment when the buyer owns the asset. Compare this with the outcome when the seller owns the asset, distinguishing between the situations where v is high and v is low. Note: for this part, assume that, under S-ownership, B's outside option is $\bar{w} < -v$, making it irrelevant.
Long Hint: this part is a bit complicated due to the non-standard bargaining game, but it is illustrative of how the bargaining structure affects investment incentives (and it makes Nash bargaining look very nice in comparison). This hint is meant to guide you through the problem.

- Under seller ownership, the bargaining game is such that the buyer chooses e_D to
$$\max_{e_D} \left\{ \min \left\{ e_D - v, \frac{e_D}{2} \right\} - \frac{c}{2} e_D^2 \right\}.$$
- Break it up into cases:
  - If $e_D - v < e_D/2$, then what is the buyer's optimal choice of e_D? Plug back in to check that the condition holds.
  - If $e_D - v > e_D/2$, then what is the buyer's optimal choice of e_D? Plug back in to check that the condition holds—what happens if it does not?
- Write the buyer's optimal choice of e_D as a step function with arguments v and c.

5.2 Foundations of Incomplete Contracts

Property rights have value when contracts are incomplete because they determine who has residual rights of control, which in turn protects that party (and its relationship-specific investments) from expropriation by its trading partners. We will now discuss some of the commonly given reasons for why contracts might be incomplete, and in particular, we will focus on whether it makes sense to apply these reasons as justifications for incomplete contracts in the Property Rights Theory.

Contracts may not be as complete as parties would like for one of three reasons. First, parties might have private information. This is the typical reason given for why, in our discussion of moral hazard models, contracts could only depend on output or a misaligned performance measure rather than directly on the agent's effort. But in such models, contracts specified in advance are likely to be just as incomplete as contracts that are filled in at a later date. We typically do not refer to such models as models of incomplete contracting models, and we reserve the term "incomplete" to refer to a contract that simply does not lay out all the future contingencies.

One often-given justification for incomplete contracts (in this more precise sense) is that it may just be costly to write a complicated state-contingent decision rule into a contract that is enforceable by a third party. This is surely important, and several authors have modeled this idea explicitly (Dye, 1985; Bajari and Tadelis, 2001; and Battigalli and Maggi, 2002) and drawn out some of its implications. Nevertheless, I will focus instead on the final reason.

The final reason often given is that parties may like to specify what to do in each state of the world in advance, but some of these states of the world are either unforeseen or indescribable by these parties. As a result, parties may leave the contract incomplete and "fill in the details" once more information has arrived. Decisions may be ex ante non-contractible but ex post contractible (and importantly for applied purposes, tractably derived by the economist as the solution to an efficient bargaining protocol), as in the Property Rights Theory.

I will focus on the third justification, providing some of the arguments given in a sequence of papers (Maskin and Tirole, 1999; Maskin and Moore, 1999; Maskin, 2002) about why this justification alone is insufficient if parties can foresee the payoff consequences of their actions, which they must if they are to accurately assess the payoff consequences of different allocations of property rights. In particular, these papers point out that there exists auxiliary mechanisms that are capable of ensuring truthful revelation of mutually known, payoff-relevant information as part of the unique subgame-perfect equilibrium. Therefore, even though payoff-relevant information may not be directly observable by a third-party enforcer, truthful revelation via the mechanism allows for indirect verification, which implies that any outcome attainable with ex ante describable states of the world is also attainable with ex ante indescribable states of the world.

This result is troubling in its implications for the Property Rights Theory. Comparing the effectiveness of second-best institutional arrangements (e.g., property-rights allocations) under incomplete contracts is moot when a mechanism exists that is capable of achieving, in this setting, first best outcomes. Here, I will provide an example of the types of mechanisms that have been proposed in the literature, and I will point out a couple of recent criticisms of these mechanisms.

An Example of a Subgame-Perfect Implementation Mechanism

I will first sketch an elemental hold-up model, and then I will show that it can be augmented with a subgame-perfect implementation mechanism that induces first-best outcomes.

Hold-Up Problem There is a Buyer (B) and a Seller (S). S can choose an effort level $e \in \{0, 1\}$ at cost ce, which determines how much B values the good that S produces. B values this good at $v = v_L + e\,(v_H - v_L)$. There are no outside sellers who can produce this good, and there is no external market on which the seller could sell his good if he produces it. Assume $(v_H - v_L)/2 < c < (v_H - v_L)$.

There are three periods:

1. S chooses e. e is commonly observed but unverifiable by a third party.

2. v is realized. v is commonly observed but unverifiable by a third party.

3. With probability $1/2$, B makes a take-it-or-leave-it offer to S, and with probability $1/2$, S makes a take-it-or-leave-it offer to B.

This game has a unique subgame-perfect equilibrium. At $t = 3$, if B gets to make the offer, B asks for S to sell him the good at price $p = 0$. If S gets to make the offer, S demands $p = v$ for the good. From period 1's perspective, the expected price that S will receive is $E[p] = v/2$, so S's effort-choice problem is

$$\max_{e\in\{0,1\}} \frac{1}{2}v_L + \frac{1}{2}e\left(v_H - v_L\right) - ce.$$

Since $\left(v_H - v_L\right)/2 < c$, S optimally chooses $e^* = 0$. In this model, ex ante effort incentives arise as a by-product of ex post bargaining, and as a result, the trade price may be insufficiently sensitive to S's effort choice to induce him to choose $e^* = 1$. This is the standard hold-up problem. Note that the assumption that v is commonly observed is largely important, because it simplifies the ex post bargaining problem.

Subgame-Perfect Implementation Mechanism While effort is not verifiable by a third-party court, public announcements can potentially be used in legal proceedings. Thus, the two parties can in principle write a contract that specifies trade as a function of announcements $\hat{v}$ made by B. If B always tells the truth, then his announcements can be used to set prices that induce S to choose $e = 1$. One way of doing this is to implement a mechanism that allows announcements to be challenged by S and to punish B any time he is challenged. If S challenges only when B has told a lie, then the threat of punishment will ensure truth telling.

The crux of the implementation problem, then, is to give S the power to challenge announcements, but to prevent "he said, she said" scenarios wherein S challenges B's announcements when he has in fact told the truth. The key insight of SPI mechanisms is to combine S's challenge with a test that B will pass if and only if he in fact told the truth.

To see how these mechanisms work, and to see how they could in principle solve the hold-up problem, let us suppose the players agree ex-ante to subject themselves to the following multi-stage mechanism.

1. B and S write a contract in which trade occurs at price $p(\hat{v})$. $p(\cdot)$ is commonly observed and verifiable by a third party.

2. S chooses e. e is commonly observed but unverifiable by a third party.

3. v is realized. v is commonly observed but unverifiable by a third party.

4. B announces $\hat{v} \in \{v_L, v_H\}$. $\hat{v}$ is commonly observed and verifiable by a third party.

5. S can challenge B's announcement or not. The challenge decision is commonly observed and verifiable by a third party. If S does not challenge the announcement, trade occurs at price $p(\hat{v})$. Otherwise, play proceeds to the next stage.

6. B pays a fine F to a third-party enforcer and is presented with a counter offer in which he can purchase the good at price $\hat{p}(\hat{v}) = \hat{v} + \varepsilon$. B's

decision to accept or reject the counter off is commonly observed and verifiable by a third party.

7. If B accepts the counter offer, then S receives F from the third-party enforcer. If B does not, then S also has to pay F to the third-party enforcer.

The game induced by this mechanism seems slightly complicated, but we can sketch out the game tree in a relatively straightforward manner.

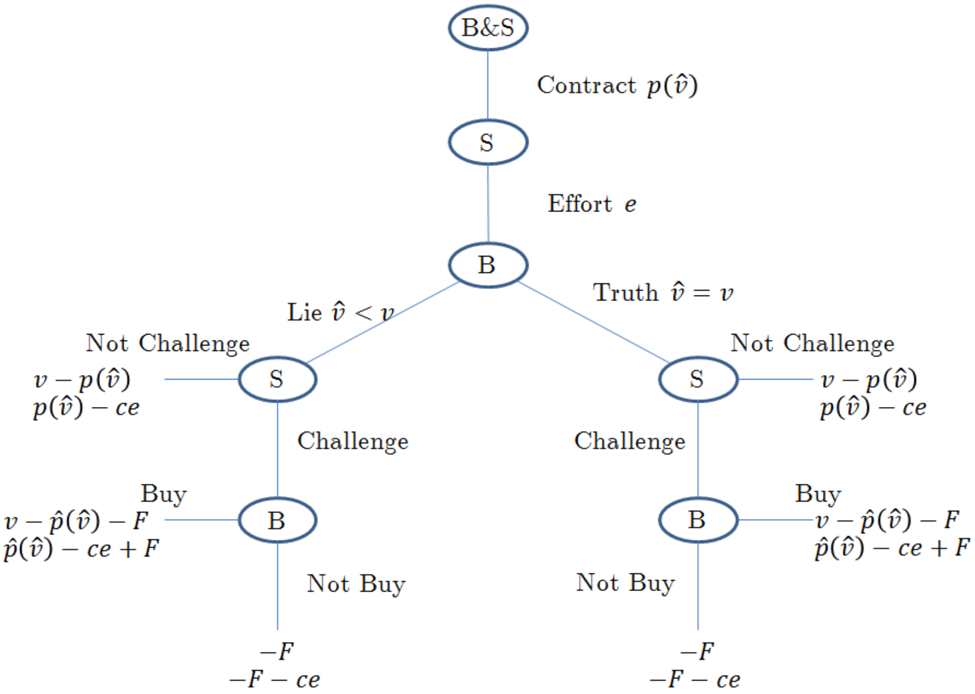


Figure 22: Maskin and Tirole mechanism

If the fine F is large enough, the unique SPNE of this game involves the following strategies. If B is challenged, he accepts the counter offer and buys the good at the counter-offer price if $\hat{v} < v$ and he rejects it if $\hat{v} \geq v$. S challenges B's announcement if and only if $\hat{v} < v$, and B announces $\hat{v} = v$. Therefore, B and S can, in the first stage, write a contract of the form $p(\hat{v}) = \hat{v} + k$, and as a result, S will choose $e^* = 1$.

To fix terminology, the mechanism starting from stage 4, after v has been realized, is a special case of the mechanisms introduced by Moore and Repullo (1988), so I will refer to that mechanism as the Moore and Repullo mechanism. The critique that messages arising from Moore and Repullo mechanisms can be used as a verifiable input into a contract to solve the hold-up problem (and indeed to implement a wide class of social choice functions) is known as the Maskin and Tirole (1999) critique. The main message of this criticism is that complete information about payoff-relevant variables and common knowledge of rationality implies that verifiability is not an important constraint to (uniquely) implement most social choice functions, including those involving efficient investments in the Property Rights Theory model.

The existence of such mechanisms is troubling for the Property Rights Theory approach. However, the limited use of implementation mechanisms in real-world environments with observable but non-verifiable information has led several recent authors to question the Maskin and Tirole critique itself. As Maskin himself asks: "To the extent that [existing institutions] do not replicate the performance of [subgame-perfect implementation mechanisms], one must ask why the market for institutions has not stepped into the breach, an important unresolved question." (Maskin, 2002, p. 728)

Recent theoretical work by Aghion et al. (2012) demonstrates that the truth-telling equilibria in Moore and Repullo mechanisms are fragile. By perturbing the information structure slightly, they show that the Moore and Repullo mechanism does not yield even approximately truthful announcements for any setting in which multi-stage mechanisms are necessary to obtain truth-telling as a unique equilibrium of an indirect mechanism. Aghion et al. (2018) takes the Moore and Repullo mechanism into the laboratory and show that indeed, when they perturb the information structure away from common knowledge of payoff-relevant variables, subjects do not make truthful announcements.

Relatedly, Fehr et al. (2017) takes an example of the entire Maskin and Tirole critique into the lab and ensure that there is common knowledge of payoff-relevant variables. They show that in the game described above, there is a strong tendency for B's to reject counter offers after they have been challenged following small lies, S's are reluctant to challenge small lies, B's tend to make announcements with $\hat{v} < v$, and S's often choose low effort levels.

These deviations from SPNE predictions are internally consistent: if indeed B's reject counter offers after being challenged for telling a small lie, then it makes sense for S to be reluctant to challenge small lies. And if S often does not challenge small lies, then it makes sense for B to lie about the value of the good. And if B is not telling the truth about the value of the good, then a contract that conditions on B's announcement may not vary sufficiently with S's effort choice to induce S to choose high effort.

The question then becomes: why do B's reject counter offers after being challenged for telling small lies if it is in their material interests to accept such counter offers? One possible explanation, which is consistent with the findings of many laboratory experiments, is that players have preferences for negative reciprocity. In particular, after B has been challenged, B must immediately pay a fine of F that he cannot recoup no matter what he does going forward. He is then asked to either accept the counter offer, in which case S is rewarded for appropriately challenging his announcement; or he can reject the counter offer (at a small, but positive, personal cost), in which case S is punished for inappropriately challenging his announcement.

The failure of subjects to play the unique SPNE of the mechanism suggests that at least one of the assumptions of Maskin and Tirole's critique is not satisfied in the lab. Since Fehr et al. (2017) is able to design the experiment to ensure common knowledge of payoff-relevant information, it must be the case that players lack common knowledge of preferences and rationality, which is also an important set of implicit assumptions that are part of Maskin and Tirole's critique. Indeed, Fehr et al. (2017) provides suggestive evidence that preferences for reciprocity are responsible for their finding that B's often reject counter offers.

The findings of Aghion et al. (2018) and Fehr et al. (2017) do not necessarily imply that it is impossible to find mechanisms in which in the unique equilibrium of the mechanisms, the hold-up problem can be effectively solved. What they do suggest, however, is that if subgame-perfect implementation mechanisms are to be more than a theoretical curiosity, they must incorporate relevant details of the environment in which they might be used. If people have preferences for reciprocity, then the mechanism should account for this. If people are concerned about whether their trading partner is rational, then the mechanism should account for this. If people are concerned that uncertainty about what their trading partner is going to do means that the mechanism imposes undue risk on them, then the mechanism should account for this.

Chapter 6

Financial Contracting

The last topic that we will cover in this class applies the tools we have developed over the last couple weeks in order to think about corporate governance, which Shleifer and Vishny (1997) define as "ways in which the suppliers of finance to corporations assure themselves of getting a return on their investment." We will think about a setting in which a capital-constrained Entrepreneur needs capital from capital-rich potential Investor to undertake a project that yields a positive return. We will look at the different instruments the Entrepreneur has to credibly commit herself to return funds to such an Investor in order to attract financing from them.

In a world of complete contracts and complete financial markets, how a project is financed—whether through debt or equity or some other, more complicated arrangement—is irrelevant for the total value of the project, and every positive net-present value project will be funded. The irrelevance result is known as the Modigliani-Miller theorem (Modigliani and Miller, 1958) and it is not so different from versions of the Coase theorem that we have mentioned in passing a few times. (Very) roughly speaking, we can think of the expected discounted revenues from the project as some value V. If undertaking the project requires K dollars worth of capital, then the Investor has to get at least K dollars back. One way he could get K dollars back is if he gets a share of the future revenues for which the expected present discounted value is K. Or the Entrepreneur could write a debt contract for which the expected present discounted value of payments is K. Either way, the Entrepreneur will receive $V-K$ and will undertake the project if $V > K$.

The Modigliani-Miller theorem served as a benchmark and spawned a literature providing explanations for when and why debt has advantages over equity based on two classes of explanations: differences in tax treatment and incentive problems. Our focus will be on the latter and in particular on how different arrangements lead the Entrepreneur to make different decisions that in turn affect the value of the project. Without appropriate contractual safeguards, the Investor might worry that the Entrepreneur will make decisions that are privately beneficial to the Entrepreneur but harmful to the Investor. The moral hazard problems that arise in these settings may include insufficient effort on the part of the Entrepreneur, although this may not take the form of the Entrepreneur working too few hours, but rather that she might avoid unpleasant tasks like firing people or a taking a tough stance in negotiations with suppliers. The problem may take the form of unnecessary or extravagant investments aimed at growing the Entrepreneur's "empire" at the expense of the Investor's returns. Or it may take the form of self dealing and excessive perk consumption: buying costly private jets, expensive art for the corporate headquarters, or hiring friends and family members.

When actions like the ones described above are not contractible, credit may be *rationed* in the sense that the Entrepreneur may be unable to "obtain the loan [she] wants even though [she] is willing to pay the interest that the lenders are asking, and perhaps even a higher interest rate." (Tirole, 2005, p. 113) Positive net-present value projects may therefore not be undertaken. We will begin with a workhorse model that builds off our analysis of limited liability constraints to provide a reason why credit may be rationed. As in our earlier discussion of such models, the Entrepreneur must be given a rent in order to provide her with incentives to take the right action. The total returns from the Entrepreneur's project net of the incentive rents the Entrepreneur must receive is what we will refer to as her *pledgeable income*. Even if the overall income from the project would be high enough to cover the Investor's capital costs, if the Entrepreneur's pledgeable income is not, she will be unable to attract funding from the Investor.

The form of the optimal contract in this model can, depending on how you look at it, be interpreted either as a debt contract or as a contract involving outside equity. But it lacks the richness of form that real financing arrangements take. In particular, when we think of equity, we typically think of a contract in which an outside Investor owns some share of a firm's profits and is also able to exercise some limited control over some of the firm's decisions. When we think of debt, we think of contracts in which the Investor is guaranteed some payments, and if the Entrepreneur does not repay the Investor, the Investor gains control over the associated assets and can then make decisions about how they are used. The model above has no notion of control rights, so it is unable to provide a compelling argument for why such contracts might move around control rights in a contingent way. We will therefore take an incomplete contracts view to think about how contingent control rights might be used in an optimal arrangement.

6.1 Pledgeable Income and Credit Rationing

There is a risk-neutral Entrepreneur (E) and a risk-neutral Investor (I). The Investor has capital but no project, and the Entrepreneur has a project but no capital. In order to pursue the project, the Entrepreneur needs K units of capital. Once the project has been pursued, the project yields revenues py, where $y \in \{0,1\}$ is the project's output, and p is the market price for that output. The Entrepreneur chooses an action $e \in [0,1]$ that determines the probability of a successful project, $\Pr[y=1|\, e] = e$, as well as a private benefit $b(e)$ that accrues to the Entrepreneur, where b is strictly decreasing and concave in e and satisfies $b'(0) = 0$ and $\lim_{e\to 1} b'(e) = -\infty$.

The Entrepreneur can write a contract $w \in \mathcal{W} = \{w : \{0,1\} \to \mathbb{R},\ 0 \leq w(y) \leq py\}$ that pays the Investor $w(y)$ if output is y and therefore shares the projects revenues with the Investor. If the Investor declines the contract, he keeps the K units of capital, and the Entrepreneur receives a payoff of 0. If the Investor accepts the contract, the Entrepreneur's and Investor's preferences are

$$\begin{aligned} U_E(w,e) &= E\left[py - w(y)\right|e] + b(e) \\ U_I(w,e) &= E\left[w(y)\right|e]. \end{aligned}$$

There are strong parallels between this model and the limited-liability Principal-Agent model we studied earlier. We can think of the Entrepreneur as the Agent and the Investor as the Principal. There is one substantive difference and two cosmetic differences. The substantive difference is that the Entrepreneur is the one writing the contract, and while the contract must still satisfy the Entrepreneur's incentive-compatibility constraint, the individual rationality constraint it has to satisfy is the *Investor's*. The two cosmetic differences are: (1) the payments in the contract flow from the Entrepreneur to the Investor, and (2) instead of higher values of e costing the Entrepreneur $c(e)$, they reduce her private benefits $b(e)$.

Timing The timing of the game is as follows.

1. E offers I a contract $w(y)$, which is commonly observed.

2. I accepts the contract $(d = 1)$ or rejects it $(d = 0)$ and keeps K, and the game ends. This decision is commonly observed.

3. If I accepts the contract, E chooses action e and receives private benefit $b(e)$. e is only observed by E.

4. Output $y \in \{0,1\}$ is drawn, with $\Pr[y=1|\, e] = e$. y is commonly observed.

5. E pays I an amount $w(y)$. This payment is commonly observed.

Equilibrium The solution concept is the same as always. A **pure-strategy subgame-perfect equilibrium** is a contract $w^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0,1\}$, an action choice $e^* : \mathcal{W} \times \{0,1\} \to [0,1]$ such that given contract w^*, the Investor optimally chooses d^*, and the Entrepreneur optimally chooses e^*, and given d^*, the Investor optimally offers contract w^*. We will say that the optimal contract induces action e^*.

The Program The Entrepreneur offers a contract $w \in \mathcal{W}$, which specifies a payment $w(0) = 0$ and $0 \leq w(1) \leq p$ and proposes an action e to solve

$$\max_{w(1),e} (p - w(1))\, e + b(e)$$

subject to the incentive-compatibility constraint

$$e \in \operatorname*{argmax}_{\hat{e} \in [0,1]} (p - w(1))\, \hat{e} + b(\hat{e}),$$

the Investor's individual-rationality (or break-even) constraint

$$w(1)\,e \geq K.$$

Analysis We can decompose the problem into two steps. First, we can ask: for a given action e, how much rents must the Entrepreneur receive in order to choose action e, and therefore, what is the maximum amount that the Investor can be promised if the Entrepreneur chooses e? Second, we can ask: given that the Investor must receive K, what action e^* maximizes the Entrepreneur's expected payoff?

The following figure illustrates the problem using a graph similar to the one we looked at when we thought about limited liability constraints. The horizontal axis is the Entrepreneur's action e, and the segment pe is the expected revenues as a function of e. The dashed line $(p - w_{e_1})\,e$ represents, for a contract that pays the Investor $w(1) = w_{e_1}$ if $y = 1$, the Entrepreneur's expected monetary payoff, and $-b(e)$ represents the Entrepreneur's cost of choosing different actions. As the figure illustrates, the contract that gets the Entrepreneur to choose action e_1 can pay the Investor at most $w_{e_1}e_1$ in expectation.

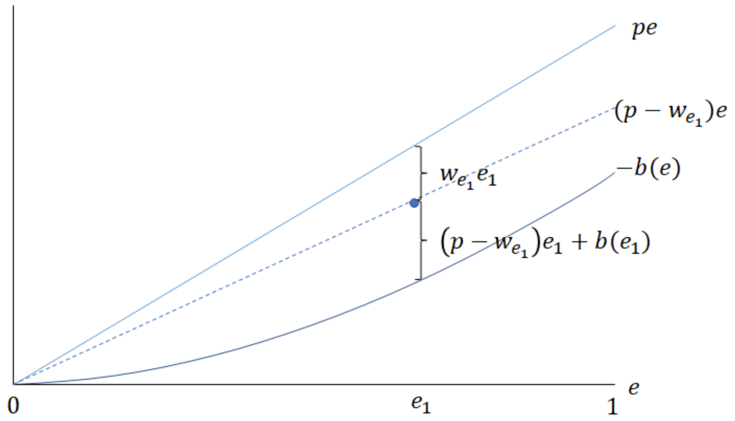


Figure 23: Entrepreneur Incentive Rents

The next figure illustrates, for different actions e, the rents $(p - w_e)\,e + b\,(e)$ that the Entrepreneur must receive for e to be incentive-compatible. Note that because $w_e \geq 0$, there is no incentive-compatible contract that gets the Entrepreneur to choose any action $e > e^{FB}$. The vertical distance between the expected revenue pe curve and the Entrepreneur rents curve is the Investor's expected payoff under the contract that gets the Entrepreneur to choose action e. For the Investor to be willing to sign such a contract, that vertical distance must be at least K, which is the amount of capital the Entrepreneur needs.

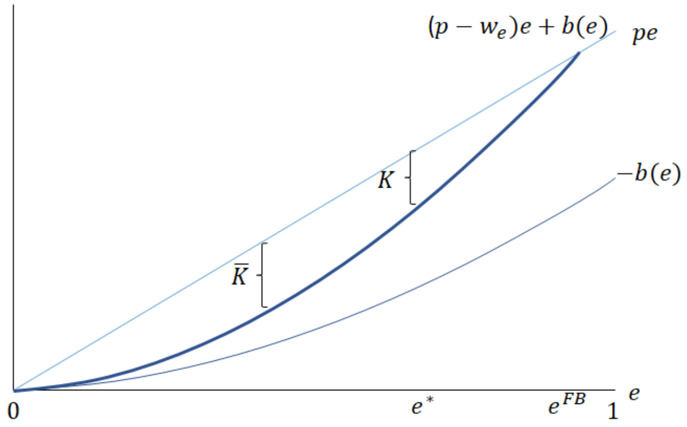


Figure 24: Equilibrium and Pledgeable Income

Two results emerge from this analysis. First, if $K > 0$, then in order to secure funding K, the Entrepreneur must share some of the project's earnings with the Investor, which means that the Entrepreneur does not receive all the returns from her actions and therefore will choose an action $e^* < e^{FB}$. Second, the value $\bar{K}$ represents the maximum expected payments the Entrepreneur can promise the Investor in any incentive-compatible contract. This value is referred to as the Entrepreneur's **pledgeable income**. If the project requires capital $K > \bar{K}$, then there is no contract the Entrepreneur can offer the Investor that the Investor will be willing to sign, even though the Entrepreneur would invest in the project if she had her own capital. When this is the case, we say that there is **credit rationing**.

As a final point about this model, with binary output, the optimal contract can be interpreted as either a debt contract or an equity contract. Under the debt contract interpretation, the Entrepreneur must reimburse w_{e^*} or else go bankrupt, and if the project is successful, she keeps the residual $p - w_{e^*}$. Under the equity contract interpretation, the Entrepreneur holds a share $(p - w_{e^*})/p$ of the project's equity, and the Investor holds a share w_{e^*}/p of the project's equity. That the optimal contract can be interpreted as either a debt contract or an equity contract highlights that if we want to actually understand the role of debt or equity contracts, we will need a richer model.

6.2 Control Rights and Financial Contracting

The previous model cannot explain the fact that equity has voting power while debt does not, except following default. Aghion and Bolton (1992) takes an incomplete contracting approach to thinking about financial contracting and brings control rights front and center. We will look at a simple version of the model that provides an explanation for debt contracts featuring *contingent* control. In this model, control rights matter because the parties disagree about important decisions that are ex ante noncontractible. The parties will renegotiate over these decisions ex post, but because the Entrepreneur is wealth-constrained, renegotiation may not fully resolve the disagreement. Investor control will therefore lead to a smaller pie ex post, but the Investor will receive a larger share of that pie. As a result, even though Investor control destroys value, it may be the only way to get the Investor to be willing to invest to begin with.

The Model As in the previous model, there is a risk-neutral Entrepreneur (E) and a risk-neutral investor (I). The Investor has capital but no project, the Entrepreneur has a project but no capital, and the project costs K. The parties enter into an agreement, which specifies who will possess the right to make a decision $d \in \mathbb{R}_+$ once that decision needs to be made. After the state $\theta \in \mathbb{R}_+$, which is drawn according to density $f(\theta)$, is realized, the decision d is made. This decision determines verifiable profits $y(d)$, which we will assume accrue to the Investor.[1] It also determines nonverifiable private benefits $b(d)$ that accrue to the Entrepreneur.

The parties can contract upon a rule that specifies who will get to make the decision d in which state of the world: let $g : \mathbb{R}_+ \to \{E, I\}$ denote the **governance structure**, where $g(\theta) \in \{E, I\}$ says who gets to make the decision d in state θ. The decision d is itself not ex ante contractible, but it is ex post contractible, so that the parties can negotiate over it ex post. In particular, we will assume that the Entrepreneur has all the bargaining power, so that she will propose a take-it-or-leave-it offer specifying a decision d as well as a transfer $w \geq 0$ from the Investor to the Entrepreneur. Note

[1] We could enrich the model to allow the parties to contract ex ante on the split of the verifiable profits that each party receives. Giving all the verifiable profits to the Investor maximizes the efficiency of the project because it maximizes the pledgeable income that he can receive without having to distort ex post decision making.

that the transfer has to be nonnegative, because the Entrepreneur is cash-constrained.

Timing

1. E proposes a governance structure g. g is commonly observed.
2. I chooses whether or not to go ahead with the investment. This decision is commonly observed.
3. The state θ is realized and is commonly observed.
4. E makes a take-it-or-leave-it offer of (d, w) to I, who either accepts or rejects it.
5. If I rejects the offer, party $g(\theta)$ chooses d.

Analysis As usual, let us start by describing the first-best decision that maximizes the sum of the profits and the private benefits:

$$d^{FB} \in \underset{d \in \mathbb{R}_+}{\operatorname{argmax}} \, y(d) + b(d).$$

Assume y and b are strictly concave and single-peaked, so that there is a unique first-best decision. Moreover, assume $y(d)$ is maximized at some decision d^I, and $b(d)$ is maximized at some other decision $d^E < d^I$. These assumptions imply that $d^E < d^{FB} < d^I$. Now, let us see what happens depending on who has control.

We will first look at what happens under Entrepreneur control. This corresponds to $g(\theta) = E$ for all θ. In this case, if the Investor rejects the Entrepreneur's offer in stage 4, the Entrepreneur will choose d to maximize her private benefit and will therefore choose d^E. Recall that the Entrepreneur does not care about the profits of the project because we have assumed that the profits accrue directly to the Investor. The decision d^E is therefore the Investor's outside option in stage 4. It will not be the decision that is actually made, however, because the Entrepreneur can offer to make a higher decision in exchange for some money. In particular, she will offer $\left(d^{FB}, w\right)$, where w is chosen to extract all the ex post surplus from the Investor:

$$y\left(d^{FB}\right) - w = y\left(d^E\right) \text{ or } w = y\left(d^{FB}\right) - y\left(d^E\right) > 0.$$

Under Entrepreneur control, the Entrepreneur's payoff will therefore be $b\left(d^{FB}\right) + y\left(d^{FB}\right) - y\left(d^E\right) > b\left(d^E\right)$, and the Investor's payoff will be $y\left(d^E\right)$, which is effectively the Entrepreneur's pledgeable income. If $y\left(d^E\right) > K$, then the Investor will make the investment, and the first-best decision will be made, but if $y\left(d^E\right) < K$, this arrangement will not get the Investor to make the investment.

Now let us look at what happens under Investor control, which corresponds to $g(\theta) = I$ for all θ. In this case, if the Investor rejects the Entrepreneur's offer at stage 4, the Investor will choose d to maximize profits and will therefore choose d^I. The decision d^I is therefore the Investor's outside option in stage 4. At stage 4, the Entrepreneur would like to get the Inventor to make a decision $d < d^I$, but in order to get him to do so, she would have to choose $w < 0$, which is not feasible. As a result, d^I will in fact be the decision that is made. Under Investor control, the Entrepreneur's payoff will be $b\left(d^I\right)$, and the Investor's payoff will be $y\left(d^I\right)$, which again is effectively the Entrepreneur's pledgeable income. Conditional on the investment being made, total surplus under Investor control is lower than under Entrepreneur control, but the benefit of Investor control is that it ensures the Investor a payoff of $y\left(d^I\right)$, which may exceed K even if $y\left(d^E\right)$ does not.

As in the Property Rights Theory, decision rights determine parties' outside options in renegotiations, which determines their incentives to make investments that are specific to the relationship. In contrast to the PRT, however, ex post renegotiation does not always lead to a surplus-maximizing outcome because the Entrepreneur is wealth-constrained. As such, in order to provide the Investor with incentives to make the relationship-specific investment of investing in the project, we may have to give the Investor ex post control, even though he will use it in a way that destroys total surplus.

If $y\left(d^I\right) > K > y\left(d^E\right)$, then Investor control is better than Entrepreneur control because it ensures the Investor will invest, but in some sense, it involves throwing away more surplus than necessary. In particular, consider a governance structure $g\left(\cdot\right)$ under which the Entrepreneur has control with probability π (i.e., $\Pr\left[g\left(\theta\right) = E\right] = \pi$), and the Investor has control with probability $1 - \pi$ (i.e., $\Pr\left[g\left(\theta\right) = I\right] = 1 - \pi$). The Entrepreneur can get the

Investor to invest if she chooses π to satisfy

$$\pi y\left(d^{E}\right)+(1-\pi)\, y\left(d^{I}\right)=K,$$

which will be optimal.

Now, stochastic control in this sense is a bit tricky to interpret, but with a slight elaboration of the model, it has a more natural interpretation. In particular, suppose that the state of the world, θ, determines how sensitive the project's profits are to the decision, so that

$$y\left(d,\theta\right)=\alpha\left(\theta\right) y\left(d\right)+\beta\left(\theta\right),$$

where $\alpha\left(\theta\right)>0$, and $\alpha'\left(\theta\right)<0$. In this case, the optimal governance structure would involve a cutoff θ^{*} so that $g\left(\theta\right)=E$ if $\theta>\theta^{*}$ and $g\left(\theta\right)=I$ if $\theta\leq\theta^{*}$, where this cutoff is chosen so that the Investor's expected payoffs would be K.

If $\alpha'\left(\theta\right) y\left(d\right)+\beta'\left(\theta\right)>0$ for all d, then high-θ states correspond to high-profit states, and this optimal arrangement looks somewhat like a debt contract that gives control to the creditor in bad states and gives control to the Entrepreneur in the good states. In this sense, the model captures an important aspect of debt contracts, namely that they involve contingent allocations of control. This theory of debt contracting is not entirely compelling, though, because the most basic feature of debt contracts is that the shift in control to the Investor occurs *only if the Entrepreneur does not make* *a repayment*. The last model we will look at will have this feature.

6.3 Cash Diversion and Liquidation

We will look at one final model that involves an important decision that is often specified in debt contracts: whether to liquidate an ongoing project. We will show that when the firm's cash flows are noncontractible, giving the Investor the rights to the proceeds from a liquidation event can protect him from short-run expropriation from an Entrepreneur who may want to direct the project's cash flows toward her own interests. The model is related to Hart and Moore's (1998) model.

The Model As before, there is a risk-neutral Entrepreneur (E) and a risk-neutral investor (I). The Investor has capital but no project, the Entrepreneur has a project but no capital, and the project costs K. If the project is funded, it yields income over two periods, which accrue to the Entrepreneur. In the first period, it produces output $y_1 \in \mathcal{Y}_1 \equiv \{0, 1\}$, where $\Pr[y_1 = 1] = q$, and that output generates a cash flow of $p_1 y_1$. After y_1 is realized, the Entrepreneur can make a cash payment $0 \leq \hat{w}_1 \leq p_1 y_1$ to the Investor. The project can then be terminated, yielding a liquidation value of L, where $0 \leq L \leq K$, which accrues to the Investor. Denote the probability the project is continued by $r \in [0, 1]$. If the project is continued, in the second period, it produces output $y_2 = 1$, and that output generates cash flow of p_2. At this point, the Entrepreneur can again make a cash payment $0 \leq \hat{w}_2 \leq p_2$ to the Investor.

The cash flows are noncontractible, so the parties are unable to write a contract that specifies output-contingent repayments from the Entrepreneur to the Investor, but they can write a contract that specifies probabilities $r : \mathbb{R}_+ \to [0,1]$ that determine the probability $r(\hat{w}_1)$ the project is continued if the Entrepreneur pays the Investor $\hat{w}_1$. The contracting space is therefore $\mathcal{W} = \{r : \mathbb{R}_+ \to [0,1]\}$. The players' payoffs, if the Investor invests K in the project are:

$$\begin{aligned} u_E(\ell, y_1, \hat{w}_1, \hat{w}_2) &= p_1 y_1 - \hat{w}_1 + r(\hat{w}_1)(p_2 - \hat{w}_2) \\ u_I(\ell, y_1, \hat{w}_1, \hat{w}_2) &= \hat{w}_1 + (1 - r(\hat{w}_1)) L + r(\hat{w}_1) \hat{w}_2. \end{aligned}$$

Throughout, we will assume that $p_2 > L$, so that liquidation strictly reduces total surplus.

Timing The timing of the game is as follows.

1. E offers I a contract $r(\hat{w}_1)$, which is commonly observed.

2. I accepts the contract $(d = 1)$ or rejects it $(d = 0)$ and keeps K, and the game ends. This decision is commonly observed.

3. If I accepts the contract, output $y_1 \in \{0, 1\}$ is realized. y_1 is commonly observed.

4. E makes a payment $0 \leq \hat{w}_1 \leq p_1 y_1$ to I. $\hat{w}_1$ is commonly observed.

5. The project is liquidated with probability $1 - r(\hat{w}_1)$. The liquidation event is commonly observed.

6. If the project has not been liquidated, output $y_2 = 1$ is realized. y_2 is commonly observed.

7. E makes a payment $0 \leq \hat{w}_2 \leq y_2$ to I. $\hat{w}_2$ is commonly observed.

Equilibrium The solution concept is the same as always. A **pure-strategy subgame-perfect equilibrium** is a continuation function $r^* \in \mathcal{W}$, an acceptance decision $d^* : \mathcal{W} \to \{0,1\}$, a first-period payment rule $w_1^* : \mathcal{W} \times \{0,1\} \to \mathbb{R}_+$, and a second-period payment rule $w_2^* : \mathcal{W} \times \{0,1\} \times \{0,1\} \times \mathbb{R}_+ \to \mathbb{R}_+$ such that given continuation function r^* and payment rules w_1^* and w_2^*, the Investor optimally chooses d^*, and given d^*, the Entrepreneur optimally offers continuation function r^* and chooses payment rules w_1^* and w_2^*.

The Program Models such as this one, in which the Entrepreneur's repayment decisions are not contractible, are referred to as **cash diversion** models. The Entrepreneur's problem will be to write a contract that specifies continuation probabilities and repayment amounts so that given those repayment-contingent continuation probabilities, the Entrepreneur will actually follow through with those repayments, and the Investor will at least break even. In this setting, it is clear that in any subgame-perfect equilibrium, the Entrepreneur will not make any positive payment $\hat{w}_2 > 0$, since she receives nothing in return for doing so. Moreover, it will be without loss of generality for the Entrepreneur to specify a single repayment amount $0 < w_1 \leq p_1$ to be repaid if $y_1 = 1$, and a pair of probabilities r_0 and r_1, where r_0 is the probability the project is continued (and not liquidated) if $\hat{w}_1 \neq w_1$, and r_1 is the probability the project is continued if $\hat{w}_1 = w_1$. The Entrepreneur's problem is therefore

$$\max_{r_0, r_1, w_1 \leq p_1} q\left(p_1 - w_1 + r_1 p_2\right) + (1 - q)\, r_0 p_2$$

subject to the Entrepreneur's incentive-compatibility constraint

$$p_1 - w_1 + r_1 p_2 \geq p_1 + r_0 p_2$$

and the Investor's break-even constraint

$$q\left(w_1 + (1 - r_1)\, L\right) + (1 - q)\,(1 - r_0)\, L \geq K.$$

It will be useful to rewrite the incentive-compatibility constraint as

$$(r_1 - r_0)\, p_2 \geq w_1,$$

which says that in order for repayment w_1 to be incentive-compatible, it has to be the case that by making the payment w_1 (instead of paying zero), the probability r_1 that the project is continued (and hence the Entrepreneur receives p_2) if she makes the payment is sufficiently high relative to the probability r_0 the project is continued when she does not make the payment.

Analysis In order to avoid multiple cases, we will assume that

$$p_1 > \frac{p_2}{qp_2 + (1-p)\,L} K,$$

which will ensure that in the optimal contract, the Entrepreneur's first-period payment will satisfy $w_1^* < p_1$.

The Entrepreneur's problem is just a constrained maximization problem with a linear objective function and linear constraints, so it can in principle be easily solved using standard linear-programming techniques. We will instead solve the problem by thinking about a few perturbations that, at the optimum, must not be profitable. Taking this approach allows us to get some intuition for why the optimal contract will take the form it does.

First, we will observe that the Investor's break-even constraint must be binding in any optimal contract. To see why, notice that if the constraint were not binding, we could reduce the payment amount w_1 by a little bit and still maintain the break-even constraint. Reducing w_1 makes the incentive-compatibility constraint easier to satisfy, and it increases the Entrepreneur's objective function. This argument tells us that the Entrepreneur will receive all of the surplus the project generates, so her problem is to maximize that surplus.

The second observation is that in any optimal contract, the project is never liquidated following repayment. To see why, suppose $r_0 < r_1 < 1$ so that the project is continued with probability less than one following repayment. Consider an alternative contract in which r_1 is increased to $r_1 + \varepsilon$, for $\varepsilon > 0$ small. Since making this change alone will violate the Investor's breakeven constraint, let us also increase w_1 by εL so that

$$w_1 + \varepsilon L + (1 - r_1 - \varepsilon) L = w_1 + (1 - r_1) L.$$

Under this perturbation, the Investor's breakeven constraint is still satisfied, and the Entrepreneur's incentive-compatibility constraint is satisfied as long as

$$(r_1 + \varepsilon - r_0) p_2 \geq w_1 + \varepsilon L,$$

which is true because $(r_1 - r_0) p_1 \geq w_1$ (or else the original contract did not satisfy IC) and $\varepsilon (p_2 - L) > 0$ since continuing the project is optimal (i.e., $p_2 > L$). If the original contract satisfied IC and IR, then so does this one, but this one also increases the Entrepreneur's objective by $q(-\varepsilon L + \varepsilon p_2)$, which again is strictly positive, since $p_2 > L$. This perturbation shows that increasing the probability of continuing the project following repayment is good for two reasons: it reduces the probability of inefficient liquidation, and it increases the Entrepreneur's incentives to repay.

Finally, the last step will be to show that the incentive constraint must bind at the optimum. It clearly must be the case that $r_0 < 1$, or else the incentive constraint would be violated. Again, suppose that the incentive constraint was not binding. Then consider a perturbation in which we raise r_0 to $r_0 + \varepsilon$, and to maintain the breakeven constraint, we increase w_1 to $w_1 + \varepsilon L (1 - q) / q$. If the incentive constraint was not binding, then it will still be satisfied if r_0 is raised by a little bit. Lastly, this perturbation increases the Entrepreneur's payoff by

$$-q \left[\frac{\varepsilon L (1 - q)}{q} \right] + (1 - q) \varepsilon p_2 = (1 - q)(p_2 - L) \varepsilon > 0.$$

In other words, if the incentive constraint is not binding, it is more efficient for the Entrepreneur to pay the Investor with cash than with an increased probability of liquidation, and since the Entrepreneur captures all the surplus, she will choose to pay in this more efficient way as much as she can.

To summarize, these three perturbations show that any optimal contract in this setting has to satisfy

$$(1 - r_0^*) p_2 = w_1^*$$

and

$$q w_1^* + (1 - q)(1 - r_0^*) L = K.$$

This is just two equations in two unknowns, so we can solve for the probability that the project is liquidated following nonpayment:

$$1 - r_0^* = \frac{K}{qp_2 + (1-q)\,L} > 0.$$

There is a complementarity between the repayment amount and the liquidation probability: if the project requires a lot of capital (i.e., K is large), then the Investor needs to be assured a bigger payment, and in order to assure that bigger payment, the project has to be liquidated with higher probability following nonpayment. If the project has high second-period cash flows (i.e., p_2 is high), then the Entrepreneur loses a lot following nonpayment, so the project does not need to be liquidated with as high of a probability to ensure repayment. Finally, if the liquidation value of the project is high, then the Investor earns more upon liquidation, so he can break even at a lower liquidation probability.

Under the first-best outcome, the project will never be liquidated, and the project will be undertaken as long as the expected cash flows exceed the required capital, or $qp_1 + p_2 > K$. The model features two sources of inefficiencies relative to the first-best outcome. First, in order to assure repayment, the Entrepreneur commits to a contract that with some probability inefficiently liquidates the project.

Second, there is credit rationing: the maximum amount the Entrepreneur can promise the Investor is p_2 in the event that output is high in the first period and L in the event that it is not, so if

$$qp_2 + (1 - q)\, L < K < qp_1 + p_2,$$

the project will be one that should be undertaken but, in equilibrium, will not be undertaken. The liquidation value of the project is related to the collateral value of the assets underlying the project, and there is a literature beginning with Kiyotaki and Moore (1997) that endogenizes the market value of those assets and shows there can be important general equilibrium spillovers across firms.

Bibliography

[1] Acemoglu, Daron. 2009. *Modern Economic Growth.* Princeton University Press.

[2] Aghion, Philippe and Patrick Bolton. 1992. An Incomplete Contracts Approach to Financial Contracting. *Review of Economic Studies*, 59(3): 473-494.

[3] Aghion, Philippe, Ernst Fehr, Richard Holden, and Tom Wilkening. 2018. The Role of Bounded Rationality and Imperfect Information in Subgame Perfect Implementation: An Empirical Investigation. *Journal of the European Economic Association.*

[4] Aghion, Philippe, Drew Fudenberg, Richard Holden, Takashi Kunimoto, and Oliver Tercieux. 2012. Subgame-Perfect Implementation Under Information Perturbations. *Quarterly Journal of Economics*, 127(4): 1843-1881.

[5] Alchian, Armen and Harold Demsetz. 1972. Production, Information Costs, and Economic Organization. *American Economic Review*, 62(5): 777-795.

[6] Antras, Pol. 2003. Firms, Contracts, and Trade Structure. *Quarterly Journal of Economics*, 118(4): 1375-1418.

[7] Arrow, Kenneth. 1964. The Role of Securities in the Optimal Allocation of Risk-Bearing. *Review of Economic Studies*, 31: 91-96.

[8] Arrow, Kenneth and Gerard Debreu. 1954. Existence of an Equilibrium for a Competitive Economy. *Econometrica*, 22(3): 265-290.

[9] Arrow, Kenneth and Frank Hahn. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

[10] Aumann, Robert. 1964. Markets with a Continuum of Traders. *Econometrica*, 32(1): 39-50.

[11] Bajari, Patrick and Steven Tadelis. Incentives versus Transaction Costs: A Theory of Procurement Contracts. *RAND Journal of Economics*, 32(3): 387-407.

[12] Baker, George. 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100(3): 598-614.

[13] Baker, George. 2002. Distortion and Risk in Optimal Incentive Contracts. *Journal of Human Resources*, 37(4): 728-751.

[14] Barron, Daniel, George Georgiadis, and Jeroen Swinkels. 2017. Optimal Contracts with a Risk-Taking Agent. Mimeo.

[15] Battigalli, Pierpaolo and Giovanni Maggi. 2002. Rigidity, Discretion, and the Costs of Writing Contracts. *American Economic Review*, 92(4): 798-817.

[16] Brown, Donald and Rosa Matzkin. 1996. Testable Restrictions on the Equilibrium Manifold. *Econometrica*, 64(6): 1249-1262.

[17] Carroll, Gabriel. 2015. Robustness and Linear Contracts. *American Economic Review*, 105(2): 536-563.

[18] Chaigneau, Pierre, Alex Edmans, and Daniel Gottlieb. Forthcoming. Does Improved Information Improve Incentives? *Journal of Financial Economics*.

[19] Coase, Ronald. 1937. The Nature of the Firm. *Economica*, 4(16): 386-405.

[20] Coase, Ronald. 1960. The Problem of Social Cost. *Journal of Law and Economics*, 3: 1-44.

[21] Debreu, Gerard. 1959. *Theory of Value*. New York: Wiley.

[22] Debreu, Gerard. 1974. Excess Demand Functions. *Journal of Mathematical Economics*, 1: 15-21.

[23] Debreu, Gerard and Herbert Scarf. 1963. A Limit Theorem on the Core of an Economy. *International Economic Review*, 4(3): 235-246.

[24] Diamond, Peter. 1967. The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty. *American Economic Review*, 57: 759-776.

[25] Diamond, Peter. 1998. Managerial Incentives: on the Near Linearity of Optimal Compensation. *Journal of Political Economy*, 106(5): 931-957.

[26] Dye, Ronald. 1985. Costly Contract Contingencies. *International Economic Review*, 26: 233-250.

[27] Fehr, Ernst, Michael Powell, and Tom Wilkening. 2017. Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms. Mimeo.

[28] Feltham, Gerald and Jim Xie. 1994. Performance Measure Congruity and Diversity in Multi-Task Principal/Agent Relations. *Accounting Review*, 69(3): 429-453.

[29] Geanakoplos, John and Heraklis Polemarchakis. 1986. Existence, Regularity and Constrained Suboptimality of Competitive Allocations when the Asset Market is Incomplete. In *Essays in Honor of K. Arrow*, vol. III, eds W. Heller and D. Starrett. Cambridge, U.K.: Cambridge University Press.

[30] Gibbons, Robert. 2010. Inside Organizations: Pricing, Politics, and Path Dependence. *Annual Review of Economics*, 2: 337-365.

[31] Grossman, Sanford and Oliver Hart. 1983. An Analysis of the Principal-Agent Problem. *Econometrica*, 51(1): 7-45.

[32] Grossman, Sanford and Oliver Hart. 1986. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, 94(4): 691-719.

[33] Hart, Oliver. 1975. On the Optimality of Equilibrium when the Market Structure is Incomplete. *Journal of Economic Theory*, 11: 418-443.

[34] Hart, Oliver. 1995. *Firms, Contracts, and Financial Structure.* Clarendon Press, Oxford.

[35] Hart, Oliver and John Moore. 1990. Property Rights and the Nature of the Firm. *Journal of Political Economy*, 98(6): 1119-1158.

[36] Hart, Oliver and John Moore. 1998. Default and Renegotiation: A Dynamic Model of Debt. *Quarterly Journal of Economics*, 113(1): 1-41.

[37] Holmstrom, Bengt. 1979. Moral Hazard and Observability. *Bell Journal of Economics*, 10(1): 74-91.

[38] Holmstrom, Bengt. 1982. Moral Hazard in Teams. *Bell Journal of Economics*, 13(2): 324-340.

[39] Holmstrom, Bengt. 1999. The Firm as a Subeconomy. *Journal of Law, Economics, and Organization*, 15(1): 74-102.

[40] Holmstrom, Bengt and Paul Milgrom. 1987. Aggregation and Linearity in the Provision of Intertemporal Incentives. *Econometrica*, 55(2): 303-328.

[41] Holmstrom, Bengt and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7: 24-52.

[42] Holmstrom, Bengt and John Roberts. 1998. *Journal of Economic Perspectives*, 12(4): 73-94.

[43] Innes, Robert. 1990. Limited Liability and Incentive Contracting with Ex-Ante Action Choices. *Journal of Economic Theory*, 52(1): 45-67.

[44] Jewitt, Ian. 1988. Justifying the First-Order Approach to Principal-Agent Problems. *Econometrica*, 56(5): 1177-1190.

[45] Jewitt, Ian, Ohad Kadan and Jeroen Swinkels. 2008. Moral Hazard with Bounded Payments. *Journal of Economic Theory*, 143: 59-82.

[46] Joskow, Paul. 1985. Vertical Integration and Long-Term Contracts: The Case of Coal-Burning Electric Generating Plants. *Journal of Law, Economics, and Organization*, 1(1): 33-80.

[47] Kerr, Steven. 1975. On the Folly of Rewarding A, While Hoping for B. *Academy of Management Journal*, 18(4): 769-783.

[48] Kiyotaki, Nobuhiro and John Moore. 1997. Credit Cycles. *Journal of Political Economy*, 105(2): 211-248.

[49] Klein, Benjamin, Robert Crawford, and Armen Alchian. 1978. Vertical Integration, Appropriable Rents, and the Competitive Contracting Process. *Journal of Law and Economics*, 21(2): 297-326.

[50] Lipsey, R.G. and Kelvin Lancaster. 1956. The General Theory of Second Best. *Review of Economic Studies*, 24(1): 11-32.

[51] Makowski, Louis and Joseph Ostroy. 1995. Appropriation and Efficiency: A Revision of the First Theorem of Welfare Economics. *American Economic Review*, 85(4): 808-827.

[52] Mantel, Rolf. 1974. On the Characterization of Aggregate Excess Demand. *Journal of Economic Theory*, 7(3): 348-353.

[53] Mas-Colell, Andreu, Michael Whinston, and Jerry Green. 1995. *Microeconomic Theory.* Oxford: Oxford University Press.

[54] Maskin, Eric. 2002. On Indescribable Contingencies and Incomplete Contracts. *European Economic Review*, 46: 725-733.

[55] Maskin, Eric and John Moore. 1999. Implementation and Renegotiation. *Review of Economic Studies*, 66(1): 39-56.

[56] MASKIN, ERIC AND JEAN TIROLE. 1999. Unforeseen Contingencies and Incomplete Contracts. *Review of Economic Studies*, 66(1): 83-114.

[57] MIRRLEES, JAMES. 1976. The Optimal Structure of Incentives and Authority within an Organization. *Bell Journal of Economics*, 7(1): 105-131.

[58] MIRRLEES, JAMES. 1999. The Theory of Moral Hazard and Unobservable Behaviour: Part I. *Review of Economic Studies*, 66(1): 3-21.

[59] MODIGLIANI, FRANCO AND MERTON MILLER. 1958. The Cost of Capital, Corporation Finance and the Theory of Investment. *American Economic Review*, 48(3): 261-297.

[60] MOORE, JOHN AND RAFAEL REPULLO. 1988. Subgame Perfect Implementation. *Econometrica*, 56(5): 1191-1220.

[61] MORONI, SOFIA AND JEROEN SWINKELS. 2014. Existence and Non-Existence in the Moral Hazard Problem. *Journal of Economic Theory*, 150: 668-682.

[62] NASH, JOHN. 1951. Non-Cooperative Games. *Annals of Mathematics*, 54(2): 286-295.

[63] OSTROY, JOSEPH. 1980. The No-Surplus Condition as a Characterization of Perfectly Competitive Equilibrium. *Journal of Economic Theory*, 22: 183-207.

[64] Rogerson, William. 1985. The First-Order Approach to Principal-Agent Problems. *Econometrica*, 53(6): 1357-1367.

[65] Scarf, Herbert. 1960. Some Examples of Global Instability of Competitive Equilibrium. *International Economic Review*, 1(3): 157-172.

[66] Shapley, Lloyd and Martin Shubik. 1977. Trade Using a Commodity as a Means of Payment. *Journal of Political Economy*, 85: 937-968.

[67] Shavell, Steven. 1979. Risk Sharing and Incentives in the Principal and Agent Relationship. *Bell Journal of Economics*, 10(1): 55-73.

[68] Shleifer, Andrei and Robert Vishny. 1997. A Survey of Corporate Governance. *Journal of Finance*, 52: 737-783.

[69] Simon, Herbert. 1991. Organizations and Markets. *Journal of Economic Perspectives*, 5(2): 25-44.

[70] Sonnenschein, Hugo. 1973. Do Walras' Identity and Continuity Characterize the Class of Community Excess Demand Functions? *Journal of Economic Theory*, 6: 345-354.

[71] Tirole, Jean. 2005. *The Theory of Corporate Finance.* Princeton, N.J.: Princeton University Press.

[72] WILLIAMSON, OLIVER. 1971. The Vertical Integration of Production: Market Failure Considerations. *American Economic Review*, 63(2): 112-123.

[73] WILLIAMSON, OLIVER. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York, NY.

[74] WILLIAMSON, OLIVER. 1979. Transaction-Cost Economics: The Governance of Contractual Relations. *Journal of Law and Economics*, 22(2): 233-261.

[75] WILLIAMSON, OLIVER. 1985. *The Economic Institutions of Capitalism*. Free Press, New York, NY.

[76] WILLIAMSON, OLIVER. 2000. The New Institutional Economics: Taking Stock, Looking Ahead. *Journal of Economic Literature*, 38: 595-613.