Intertemporal Decision.pdf

Lesson 1.pdf

Lesson 2.pdf

Lesson 3.pdf

Lesson 4.pdf

Lesson 5.pdf

Lesson 6.pdf

Lesson 7.pdf

Lesson 8.pdf

Lesson 9.pdf

9. Intertemporal Equilibrium Theory

Not covered in Varian, although Chs 17, 19 and 20 have tangentially related material.

Good references include:

Price Theory and Applications, 6th Ed, Ch 14, by J. Hirshleifer

Financial Theory and Corporate Policy, 3rd Ed, Ch 1-2, by T. Copeland.

I. Overview

A. These notes cover the **general equilibrium** theory of production and exchange **across time**,

1. originally developed by Irving Fisher (top American economist) 100 years ago.
2. Sorts out many of the fundamentals of finance – investment, savings, real vs nominal interest rates, productivity, thrift, etc. – in a simple abstract setting.
3. (The same abstract GE setting can be re-interpreted to explain the basics of international trade!)

B. Finance is that branch of economics that deals with risky trades over time. As such, it has three pillars:

1. Choice and market trades over time – our focus in these notes.
2. Static choice and markets for risky assets – the focus of most MBA-oriented finance texts.
3. Markets for information, typically asymmetric – the focus of ongoing academic research.

C. Deep understanding of the fundamentals is essential when conditions change and established rules-of-thumb become questionable.

D. Most MBA-oriented textbooks skip the first pillar because it is hard for non-economists to understand, and they focus instead on the second pillar.

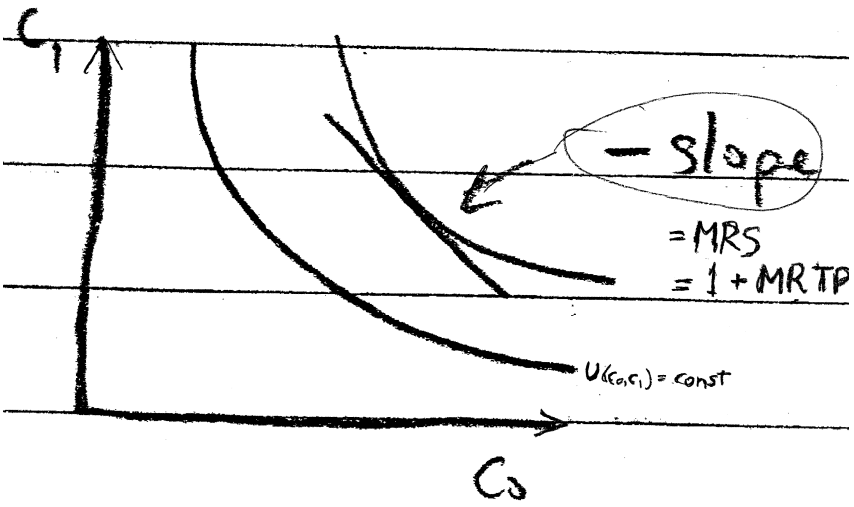


Figure 1:

II. Begin with a two-date barter model, later extend to many dates and monetary exchange.

A. There are two dates: $t = 0$ ("now") and $t = 1$ ("later").

B. There is only one good, c ("corn" or "consumption baskets").

C. There are N agents with given preferences and production opportunities, all with access to a financial market. We now specify these elements.

III. **Preferences** of agent i (index suppressed) are represented by a utility function $U(c_0, c_1)$.

A. Assume classical properties – smooth, strictly monotone, concave, Inada.

B. Inada implies that some consumption at each date is essential, and monotone implies that more consumption is always better.

$$-\text{slope of IC} = \frac{\partial_0 U}{\partial_1 U} = MRS_{01} = 1 + MRTP. \tag{1}$$

C. Most of equation (1) is familiar from an earlier unit. I'm using the shorthand notation $\partial_t U = \frac{\partial U}{\partial c_t}$ for the marginal utilities from current and future consumption, $t = 0, 1$. See Figure 1.

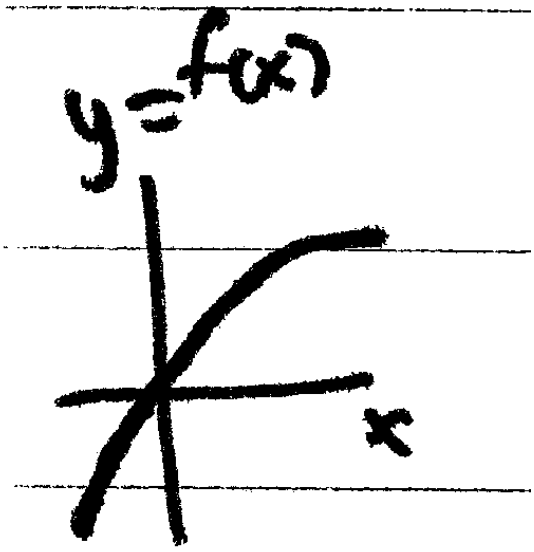


Figure 2: The production function

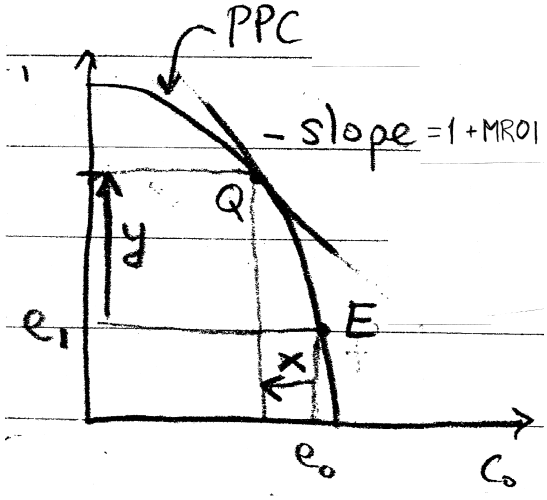


Figure 3: The PPF obtained from that production function by mapping the Origin to E and reversing the x-axis.

D. The last expression in (1) defines MRTP, the marginal rate of time preference. Equivalently, MRTP $= \frac{\partial_0 U}{\partial_1 U} - 1$.

E. That is, on the margin, agent i needs MRTP *extra* units of future consumption per unit of foregone current consumption. This turns out to be a more convenient way to express the tradeoff than MRS, which is the *total* number of units of future consumption (the return of the original unit plus the MRTP) required per unit of foregone current consumption.

IV. **Productive opportunities** of agent i (index still suppressed) are represented by a

A. production function $y = f(x)$ as in Figure 2 that relates

B. increments of future consumption $y = \Delta c_1$ to

C. the amount of foregone current consumption $x = -\Delta c_0$.

D. Think of production as planting seed corn, or more generally think about allocating overall resources between immediate consumption and building for the future.

E. As shown in Figure 3, the increments are taken relative to a given endowment point $E = (e_0, e_1)$.

F. We make the standard assumptions that $f(0) = 0, f' > 0, f'' < 0$.

G. The figure shows the corresponding production possibility frontier,
PPF $= \{(q_0, q_1) : \ q_0 = e_0 - x \geq 0, \ q_1 = e_1 + f(x) \geq 0\}$.

Handy summary descriptions are average and marginal return on investment:

$$ROI = f(x) - x, \ AROI = \frac{f(x)}{x} - 1, \ MROI = f'(x) - 1. \tag{2}$$

Or, in more geometrical terms,

$$-\text{slope of PPF} = MRT = 1 + MROI = f'(x). \tag{3}$$

V. A **financial market** is available to all agents at date $t = 0$, in which

A. units of current consumption c_0 can be traded for

B. claims (i.e., **promises**) to units of future consumption c_1.

C. The -slope of the budget line is the price ratio, which we write out as

$$-slope = \frac{p_0}{p_1} = 1 + r. \tag{4}$$

D. Presumably the price p_1 of a promise to 1 unit of c next period is less than the price p_0 of 1 unit of c held now.

E. Thus the price ratio exceeds 1.0. The degree of excess, $r \geq 0$, is called the **real interest rate**.

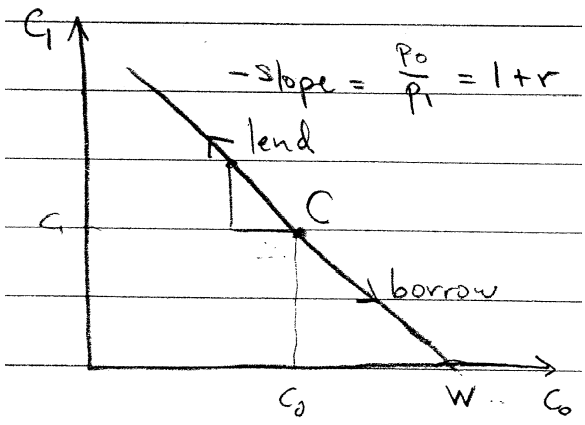


Figure 4: A trader who brings bundle $C = (c_0, c_1)$ to the financial market can trade to obtain any bundle along the budget line. The trader is borrowing [lending] if the post-trade holding of current consumption [future consumption] exceeds the pretrade value.

F. As shown in Figure 4, each unit of c_0 exchanges for $1+r$ units of c_1 in the financial market. To summarize in a diagram,

$$c_0 \xrightarrow{1+r} c_1$$

. .

VI. Results: Wealth and Present Value

A. Given a real interest rate $r > 0$ and consumption stream $C = (c_0, c_1)$, define the **present value** of C as the horizontal axis intercept w of the budget line thru C.

B. High school geometry enables us to compute $w = PV_r(C)$.

- Consider the right triangle with base $[c_0, w]$ on horizontal axis and vertex C.
- We need to compute the length z of the triangle's base in terms of the height c_1 and rise / run = -slope $= 1 + r$.
- Thus $1 + r = \frac{c_1}{z}$, so $z = \frac{c_1}{1+r}$.
- We conclude that $w = PV_r(C) = c_0 + z = c_0 + \frac{c_1}{1+r}$.

C. The intuition is that the present value is the amount of maximum amount of present consumption we can get in the financial market. It simply adds the amount c_0 of current consumption we already have to the amount $z = \frac{c_1}{1+r}$ we can get from the future part of our consumption stream.

D. In the static context, the horizontal intercept of the budget line (normalizing that price to 1) is called income. In our intertemporal context, that intercept is instead called **wealth**.

E. Hence we write $w = PV_r(C) = c_0 + \frac{c_1}{1+r}$.

VII. The next result characterizes the agent's optimum investment, given her production function f, her endowment $E = (e_0, e_1)$ and the real interest rate r.

- For reasons elaborated below, an individual agent seeks to maximize wealth by choosing the amount x of first period consumption to invest.
- Formally, her choice problem is

$$\max_x w = PV_r(Q) = q_0 + \frac{q_1}{1+r} = e_0 - x + \frac{e_1 + f(x)}{1+r}. \tag{5}$$

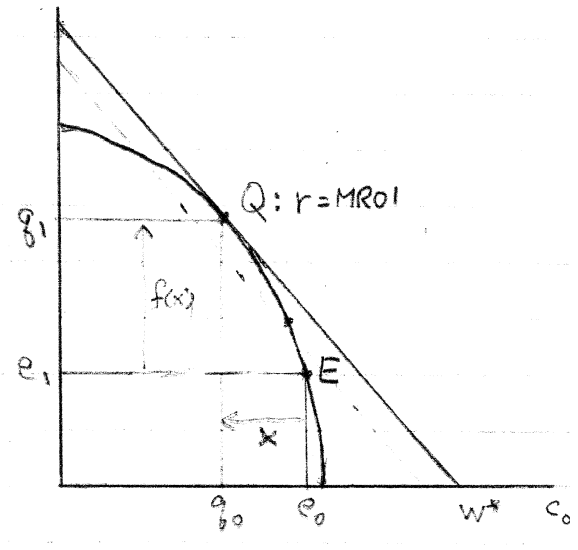


Figure 5: The optimal investment point Q is chosen along the PPF to maximize the budget line intercept w^*.

- The FOC is

$$0 = \frac{dw}{dx} = -1 + \frac{f'(x)}{1+r}. \tag{6}$$

$$\implies 1 + r = f'(x) = 1 + MROI \tag{7}$$

$$\implies r = MROI. \tag{8}$$

- That is, if not a corner solution, optimal investment is characterized by marginal return on investment equal to the real interest rate in the financial market.
- The intuition can be gleaned from Figure 5: If $MROI < r$, then $w = PV_r(Q)$ increases (i.e., the budget line thru Q shifts out) if you slightly **reduce** investment x, and if $MROI > r$, then w increases if you slightly **increase** investment x.

VIII. Fisher separation theorem

A. You might think at first that the optimal production point Q depends on the agent's preferences.

- That would be true if there were no financial market ("autarky"), but

- a financial market allows the agent to reshape her consumption stream as she pleases.
- By increasing w, she can consume more at every date.
- Thus as long as her preferences are monotone, she will maximize utility by maximizing $w = PV_r(Q)$.

B. Irving Fisher was the first to point this out clearly: optimal investment and production DOES NOT depend on preferences, as long as they are monotone. Given access to a financial market, optimal investment depends ONLY ON endowment and productive opportunities.

C. This "separation" result now seems trivial to obtain, but it still has far-reaching implications.

D. Suppose, for example, several different people own shares of a productive opportunity, and some owners are very patient (small MRTP) while others are impatient (large MRTP).

- According to the Fisher separation result, there is no need to quarrel.
- They should invest up to the point where MROI=r to maximize the present value w and thus maximize each of their shares $\alpha_i w$.
- Then each of them can borrow or lend their wealth (including that $\alpha_i w$) as they prefer.
- This is a crucial reason why corporations can exist, and why their primary goal is to maximize shareholder value.

E. editorial comments, not relevant to intertemporal equilibrium theory:

- In my opinion, a primary goal of government is to make and enforce laws that harmonize the public interest with maximizing shareholder value.
- Sometimes firms profess other goals such as serving customers, workers or the environment. These other goals may (and should) be consistent with maximizing shareholder value in the long run.

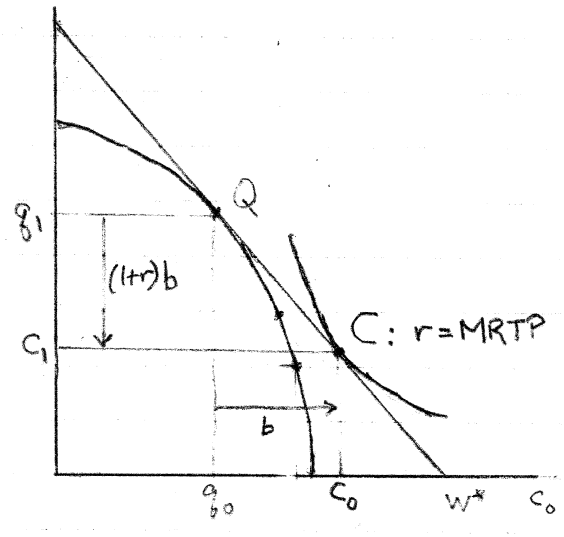


Figure 6: The optimal investment point Q is chosen along the PPF to maximize the budget line intercept w^*.

IX. Optimal individual borrowing, lending and consumption

A. Assuming that the agent has indeed maximized w, how does she choose her consumption stream? That is, how does she borrow and lend in the financial market to achieve optimal consumption over time?

B. More precisely, given the real interest rate r and her chosen production stream Q, how does she maximize utility $U(c_0, c_1)$?

C. Write her problem as

$$\max_{c_0,c_1 \geq 0} U(c_0, c_1) \text{ s.t. } PV_r(C) = PV_r(Q) = w \tag{9}$$

D. In terms of the amount $b = c_0 - q_0$ to borrow in the financial market (see Figure 6), her problem can be rewitten as just:

$$\max_b U(q_0 + b, q_1 - (1 + r)b) \tag{10}$$

E. The FOC for problem (10) (which is necessary and sufficient given our strong

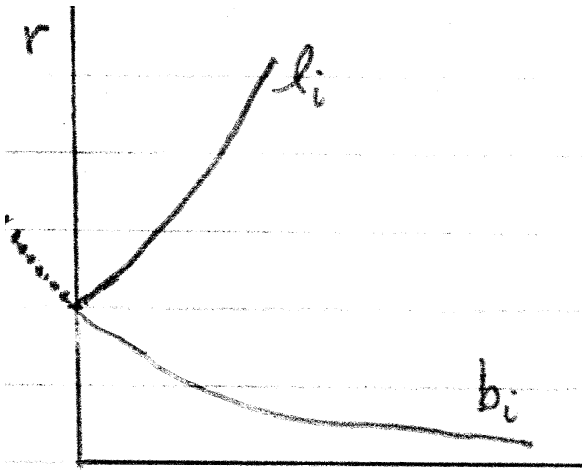


Figure 7: Amount borrowed ($b \geq 0$) or lent ($\ell \geq 0$) by agent i as functions of the real interest rate r.

classical assumptions on U) is:

$$\begin{aligned} 0 &= \partial_0 U \cdot \frac{d(q_0 + b)}{db} + \partial_1 U \cdot \frac{d(q_1 - (1+r)b)}{db} = \partial_0 U - \partial_1 U \cdot (1+r) \\ \implies (1+r) &= \frac{\partial_0 U}{\partial_1 U} = MRS = 1 + MRTP \\ \implies r &= MRTP. \end{aligned} \tag{11}$$

F. That is, optimal consumption is characterized by marginal rate of time preference equal to the real interest rate in the financial market.

G. The intuition can be gleaned from Figure 6: If $MRTP < r$ at some point along the budget line, then utility increases if you slightly **reduce** borrowing b, and if $MRTP > r$, then utility increases if you slightly **increase** b.

H. Borrowing is positive in Figure 6, but it is easy to imagine (with steeper ICs, corresponding to greater impatience) that the tangency of the IC to the budget line could occur at a point C above Q instead of below, i.e., we could have $b < 0$.

I. In this case, $-b = \ell > 0$ is called lending.

X. Equilibrium real interest rate

A. Consider the impact of a change in the real interest rate r on optimal production Q and consumption C, and hence on borrowing b or lending ℓ.

B. An increase in r, i.e., a steeper budget line, will rotate the tangency point Q on the PPF clockwise, i.e., will increase q_0 and thereby tend to reduce b. This production effect follows from the concavity of the production function ($f'' < 0$).

C. The tangency point C on the indifference curve will, by the substitution effect, also rotate clockwise, i.e., will decrease c_0 and thereby also tend to reduce b. This follows from convex preferences.

D. There is also an income effect, since the new budget line will intersect a different indifference curve. If c_0 and c_1 are both normal goods, then this income effect will again reduce b for borrowers, but will decrease ℓ for lenders.

E. This income effect is the reason why the curves drawn in Figure bl1 get steeper at higher r.

F. The Figure also shows that there is some interest rate at which $Q = C$ and so $b = \ell = 0$. I like to call this the Polonius interest rate (after the Shakespeare character who said "neither a borrower nor a lender be") but it could also be called the autarky interest rate.

G. Recall that there are $N > 1$ agents who participate in the financial market. Although we assume that they are all price-takers (i.e., each has negligible influence on r), their combined desires to borrow and lend determine the equilibrium real interest rate r^*.

H. Let $b_i(r) = \max\{0, b\}$ be agent i's borrowing curve as just derived, and $\ell_i(r) = \max\{0, \ell\}$ be her lending curve; they intersect the vertical axis and each other at her Polonius interest rate. The left panel of Figure 8 shows examples for $i = 1, 2$ where agent 1 has the lower Polonius interest rate, due to lower MROI and/or lower MRTP.

I. Aggregate borrowing is $B(r) = \sum_{i=1}^{n} b_i(r)$ and aggregate lending is $L(r) =$

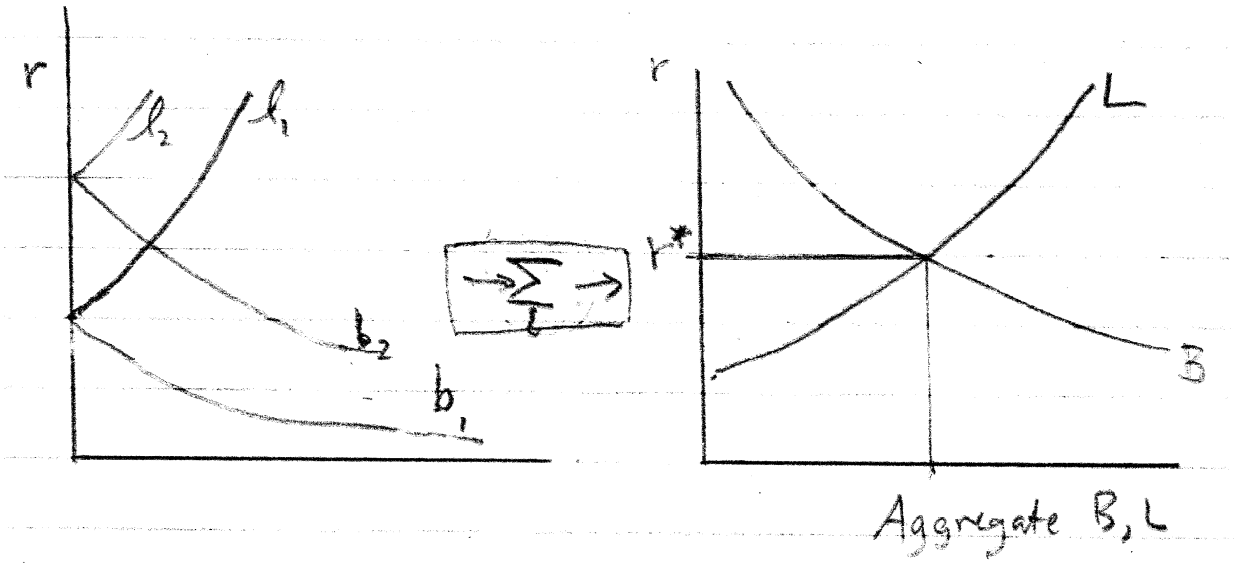


Figure 8: The equilibrium real interest rate r^* equates aggregate demand for loans to aggregate supply.

$\sum_{i=1}^{n} \ell_i(r)$. As long as agents have different Polonius r's, these curves will intersect at a positive value $B(r^*) = L(r^*)$ of financial activity.

J. Financial markets thus clear at a equilibrium real interest rate r^*. It is unique given our strong classical assumptions on production and utility functions.

K. Economists had long debated whether productivity or thrift was the main determinant of interest rates. Fisher showed how they work together to determine the (risk-free) equilibrium real interest rate.

Ex A: The transcontinental railroad unites California's financial market with the eastern US. (Hint: combine B, L).

Ex B: Info Tech increases productivity. (Hint: Use rep. agent, see impact on Polonius r)

Ex C: Aging population in developed countries increases thrift.

XI. Extending the basic model: Money

A. Still assume just 2 dates and no uncertainty, but add a second good m called

money.

1. Money is the numeraire and is storable. Its definitive role as the medium of exchange will emerge in the next extension.
2. The nominal financial market allows exchange at interest rate k, and the money price for consumption (i.e., the price level) in period $t = 0, 1$ is P_t.
3. Using notation similar to that already introduced for the real financial market, we summarize all four competitive markets in the diagram

$$\begin{array}{ccc} c_0 & \xrightarrow{P_0} & m_0 \\ {\scriptstyle 1+r}\downarrow & & \downarrow{\scriptstyle 1+k} \\ c_1 & \xrightarrow[P_1]{} & m_1 \end{array}$$

4. The bottom edge says that there is a competitive market in which you can sell each unit of c_1 for P_1 units of money, or buy $\frac{1}{P_1}$ units of c_1 for each unit of money. These forward market transactions are promises made at time $t = 0$ to be carried out at $t = 1$.
5. The top edge, of course, is for transactions carried out now ($t = 0$) in the spot market, and the left and right edges are intertemporal transactions.
6. Write $\frac{P_1}{P_0} = 1 + \pi$, where π is called the **inflation rate**, the rate of change in the price level.

B. Notice that there are two ways to go from consumption (or corn) now to money later:

   (a) along the top and left edges of the diagram (cashing out and lending), yielding $P_0 \cdot (1 + k)$ units of money at $t = 1$ per unit of c_0, or

   (b) along the right and bottom edges of the diagram (real lending and cash out the repayment), yielding $(1 + r) \cdot P_1$ units of money at $t = 1$.

C. Arbitrage ensures that both ways yield the same amount of money at $t = 1$.

- If not, say $P_0 \cdot (1 + k) > (1 + r) \cdot P_1$.

- Then go around the rectangle clockwise, doing (b) backwards and (a) frontwards, i.e., cash out, lend money, buy c_1 (in the forward market) and borrow in the "real" financial market.
- For each unit of c_0 you start with, at the end of the round trip you have

$$P_0 \cdot (1+k) \cdot \frac{1}{P_1} \cdot \frac{1}{(1+r)} = \frac{P_0 \cdot (1+k)}{(1+r) \cdot P_1} > 1$$

  units of c_0.
- You (and everyone else) can keep doing this, accumulating more and more c_0 for free, until prices adjust.
- This arbitrage buying and selling will push down P_0 and k, and push up P_1 and r. Similarly, if prices are out of line the other way, the reverse pressures will be felt.
- Hence general (multimarket) equilibrium is only possibly when

$$P_0 \cdot (1+k) = (1+r) \cdot P_1. \tag{12}$$

D. Rewrite equation (12) as $(1+k) = (1+r)\frac{P_1}{P_0}$ and recall that the last factor is $1+\pi$. Expanding this expression, we get

$$\begin{aligned} 1+k &= (1+r)(1+\pi) \\ k &= r+\pi+r\pi \approx r+\pi \end{aligned} \tag{13}$$

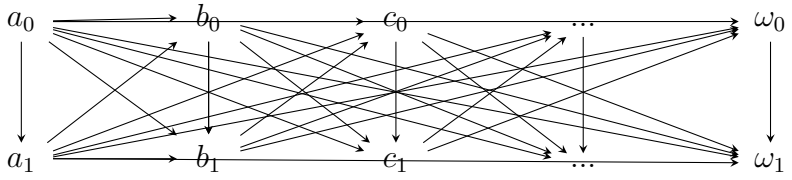
E. Equation (13) is called **Fisher's equation**. It says that the (k)nominal interest rate is the sum of the real interest rate r and the inflation rate π...

- plus a cross-term which is very small if r and π are both moderate.
- The cross-term is 0 if you use continuously compounded interest and inflation rates.
- Here is a tangent (not needed for Econ 200, but possibly handy later) on continuous compounding (or growth rates).

1. Over time interval $\Delta t > 0$, prices increase by factor $e^{\pi \Delta t}$ and real balances increase by factor $e^{r \Delta t}$.
2. Hence nominal balances rise by factor $e^{k \Delta t} = e^{r \Delta t} e^{\pi \Delta t} = e^{(r+\pi)\Delta t}$.
3. Taking logs and canceling the common factor $\Delta t > 0$, we see that with continuous compounding $k = r + \pi$.

XII. Extending the basic model: Many goods and transactions costs

A. The number of possible markets increases quadratically in the number of goods x dates. For example, with just two dates but goods $a, b, c, ...\omega$ a few of the possible markets are as in the diagram.

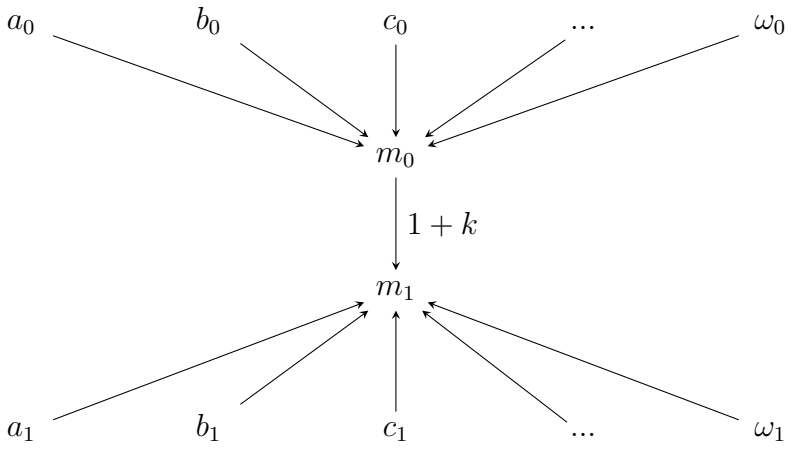


B. Even with just 1000 goods, there are $1000 \cdot 999/2 \approx 500,000$ different spot markets, plus the same number of forward markets, plus 1 million intertemporal markets, almost 2 million all together.

C. Even if the fixed and marginal costs are small for maintaining a market to trade one good against another, the logistics still get out of hand.

D. The universal solution is to abandon direct exchange ("barter") in favor of indirect exchange, using some particular good (called money) to mediate.

1. E.g., to trade a_0 for ω_0, first sell a_0 for m_0 and then use that to buy ω_0.
2. This only requires 1000 spot markets, instead of half a million.
3. Ditto for forward markets.
4. Even better: only **one** intertemporal market is needed, the nominal financial market discussed earlier.



E. Another way to put it: arbitrage renders most barter markets redundant, and they shrivel, saving transactions costs.

- In the example with 1000 goods and two dates, about 99.9% of transactions costs are saved; in realistic examples the savings are much more substantial.
- Even in a virtual economy designed to support barter, monetary exchange emerged spontaneously! [see Baumer and Kephart working paper].

F. To return to the main point, the "real" financial market does not exist in a multi-good world such as ours, only the nominal financial market. But the previous analysis lays bare the underlying forces of productivity and thrift.

XIII. Extending the basic model: Many periods

A. Suppose that money can be exchanged (borrowed and lent) for many periods T at a consistent per-period nominal interest rate k.

$$m_0 \xrightarrow{1+k} m_1 \xrightarrow{1+k} m_2 \xrightarrow{1+k} m_3 \xrightarrow{1+k} \ldots \xrightarrow{1+k} m_T$$

B. Then the $t = 1$ value of amount x of money held at date $t = 2$ is $\frac{x}{1+k}$ and its date $t = 0$ value (i.e., its present value) is $\frac{\frac{x}{1+k}}{1+k} = \frac{x}{(1+k)^2}$.

C. The diagram thus tells us that the present value of an amount x of money held at date t is $\frac{x}{(1+k)^t}$.

D. We conclude that the present value of an arbitrary money stream $X = (x_0, x_1, x_2, ..., x_T)$ is

$$PV_k(X) = \sum_{t=0}^{T} \frac{x_t}{(1+k)^t}. \tag{14}$$

- The formula applies even for an infinite horizon, $T = \infty$.
- In continuous time, the formula is

$$PV_k(X) = \int_{t=0}^{T} x_t e^{-kt} dt. \tag{15}$$

E. This brings us to the very important **intertemporal decision rule**:

If you have access to a financial market with nominal borrowing and lending rate k, then (no matter what your underlying preferences), you should strictly prefer cash stream $X = (x_0, x_1, x_2, ..., x_T)$ to cash stream $Y = (y_0, y_1, y_2, ..., y_T)$ if and only if $PV_k(X) > PV_k(Y)$.

- As before, the logic is that this choice makes your opportunity set as large as possible.
- What you choose to do with those financial market opportunities, by reshaping the cash stream to meet consumption needs or whatever else, of course does depend on your personal preferences.

F. What if the interest rate k differs from period to period?

- An arbitrage diagram similar to that above establishes the connections between one period rates for spot and forward trades (as above) and longer maturity spot rates.
- The graph of annualized spot rates k_t at all maturities is called the yield curve.
- The intertemporal decision rule still holds, except that the interest rates k_t in the PV formula come from the yield curve.

G. One last point about present value: it is a linear operator, i.e.,

$PV_k(X + Y) = PV_k(X) + PV_k(Y)$.

Financial economists refer to this property as value additivity, and use it to slice and dice financial assets.

XIV. A general formula for interest rates and asset yields:

$$k_a = r^* + \pi^e + RP_a \pm T_a \tag{16}$$

- Equation (16) says that any asset a has a yield k_a with four components.
- The first two components are the real interest rate r^* and the anticipated inflation rate π^e. By Fisher's equation, these are the only components of the risk-free nominal interest rate.
- Each asset a also has its own risk premium RP_a. Your finance course will feature theories (and evidence) on how the risk premium is determined in worlds where promises can be broken and cash streams are uncertain.
- There are also transactions costs T_a specific to asset a. These include tax; e.g., yields are higher on corporate bonds in the US than municipal bonds largely because corporate bonds have a higher tax rate. Use $+T_a$ for borrowing and $-T_a$ for lending.
- Equation (16) implies that something that increases inflation expectations (say) by 1% will push up **all** interest rates and asset yields by **exactly** 1%, assuming that that something doesn't have a separate impact on the other components. Likewise, something (like an aging workforce) that decreases the real interest rate by 2% will depress all yields by that same amount.

XV. Basic asset pricing formula:

$$P_a = PV_k(Y - C) = PV_k(Y) - PV_k(C) \tag{17}$$

- Equation (17) says that the price of any asset a is determined by its associated revenue stream $Y = (y_0, y_1, y_2, ..., y_T)$ and cost stream $C = (c_0, c_1, c_2, ..., c_T)$.

- If a market exists for the asset, then rational investors will bid up the price P_a whenever it falls below $PV_k(Y - C)$, according to the intertemporal decision rule. Likewise, they will sell off any asset with price above the **fundamental value** $PV_k(Y - C)$.
- Many financial assets do not have a competitive market, or (in some cases) any market at all. In that case, equation (17) can be used to impute a value. This is what armies of financial analysts do all day long.
- Note that P_a depends on the interest rate $k = k_a$ used in the PV formula. For most assets, the revenue and cost streams are such that an increase in k decreases P_a. That is, asset prices and yields tend to move in opposite directions.

Econ 200 Lecture Notes

by Dan Friedman, UCSC, Fall 2016

Preface These notes are are intended to help students follow in-class lectures, to organize readings, and to anticipate questions to ask in class. They mention but do not develop crucial material. They are not a substitute for reading the text!

1. Competitive Markets

The Economic Environment

I. Simplifying assumptions are essential in economic analysis. They clear the underbrush so that we can see the essential forces clearly.

II. Traditional assumptions of competitive analysis:

   A. Lots of small buyers and sellers

   B. with a whole lot of information about one another

   C. all selling the same product

   D. with no barriers to their activity.

We don't actually need assumptions B or C or even much of A, as you will soon see.

Behavioral assumption: everyone takes price as given, and doesn't try to change it.

Competitive markets $\iff$ **price-taking**.

Q: But then who sets price?

A: "impersonal forces" of supply and demand.

Exercise: Which of the above assumptions seem reasonable for the Santa Cruz apartment rental market? For mobile data plans? ...

Demo: Double Auction Market. Buyers' values and Seller's costs randomly assigned for single indivisible units. Compute market demand.

III. Market Demand Curve

A. A schedule showing the number of units buyers are willing and able to purchase as a function of the unit price.

1. This in turn is just the sum of the individual demand curves from everyone in the market.
2. Controls for "everything else" that might shift the curve...

B. Usually we assume there are lots of units demanded and so we represent demand as a continuous function.

Ex: Linear Demand:

$$q^d = a - bp \tag{1}$$

where $a, b > 0$ and equation (1) is valid for $p \in [0, a/b]$.

Ex: Log-linear demand:

$$\ln(q^d) = \ln(a) + \epsilon_d \ln(p), \quad \text{or} \quad q^d = e^a p^{\epsilon_d}. \tag{2}$$

The curve is hyperbolic and downward sloping if $\epsilon_d < 0$.

Ex: Stair step demand for indivisible units.

Ex: Inverse demand (linear):

$$p^d = \alpha - \beta q, \tag{3}$$

where, in terms of the parameters in equation (1), we have $\alpha = a/b, \beta = 1/b$.

IV. Market Supply Curve

A. A schedule showing the number of units offered by the market for sale as a function of the price charged.

   1. This in turns is just the sum of the individual supply curves from everyone in the market.
   2. Individual supply curves are actually the marginal cost curve of a firm (later we will see that no firm will supply at prices below their average variable cost).

   Ex: Supply in the Discrete [or indivisible] Goods Model

B. We usually end up assuming lots of firms and represent supply as a continuous function as well.

   Ex: Linear Supply: $q^s = A + Bp$
   Ex: Log-linear supply: $ln(q^s) = \gamma + \epsilon_s ln(p)$.
   Curves slope upward if the coefficients (B, ϵ_s) are positive.

Competitive Equilibrium

I. The next task of economic analysis is to find the consequences of the assumptions, to make predictions and draw insights. This called **positive theory**.

II. Competitive equilibrium theory predicts the quantity traded and the price in a market.

III. Intuition:

   A. If prices were too high, then there would be more units offered than demanded ($q^s > q^d$). Unsatisfied suppliers lower their asking price, and other suppliers have to match.

B. prices too low $\implies$ more units demanded than offered, ($q^d > q^s$). Unsatisfied demanders bid the price up.

Ex: Equilibrating forces in the Double Auction Market

IV. The market hopefully will quickly settle down so that there is neither excess supply nor excess demand, and equilibrium is achieved (or approximated).

V. A **competitive equilibrium** (CE) is a price p^* and resulting quantity traded q^* such that $q^s(p^*) = q^d(p^*) = q^*$.

Ex: Linear S, D.

Existence and Uniqueness

I. Is there a CE? If so, is it unique?

II. If not, the prediction is less useful !

Theorem. If

A. q^d is continuous and (strictly) decreasing in p,

B. q^s is continuous and (strictly) increasing in p,

C. $q^d(0) \geq q^s(0)$ and $q^d(\infty) \leq q^s(\infty)$

Then there is a CE (p^*, q^*) (and it is unique).

proof sketch: Apply the intermediate value theorem to excess demand function $Z(p) = q^d(p) - q^s(p)$, to find (unique) root $p^* = Z^{-1}(0)$.

Ex: Equilibrium in discrete unit model....often p^* is unique but not q^*, or vice versa.

Ex: Non-existent markets, when C3. fails.

Welfare Economics

I. Another task in economic analysis is judging an outcome as (relatively) good or bad. This is called **normative theory** , in contrast to positive theory. It also is referred to as **welfare economics**.

II. Economists main criterion is efficiency.

   A. So what is efficiency, exactly?

   B. Several versions (see Varian Ch 10 for a start) but here we'll focus on a vanilla version, Kaldor-Hicks efficiency or cost benefit analysis.

      1. Producer surplus (PS): The amount a producer is paid that is in excess of [variable] cost.
      2. Consumer surplus (CS): The amount a consumer would have paid [WTP, on the demand curve] in excess of the amount she actually had to pay.
      3. Total surplus: TS = CS + PS
      4. Efficiency = [normalized] TS.

Ex: Surplus in the Discrete Goods Model

III. Efficiency sounds cold and clinical but economists absolutely love it, as our highest virtue. And rightly so: increasing efficiency means either directly making more people happier, or conserving resources (which then can be used to increase happiness.)

IV. Economists sometimes call TS the gains from trade

   A. It is a measure of the amount of happiness generated purely by the act of trading.

V. On typical supply and demand graphs, total surplus is the area to the left of the number of units produced which is under the demand curve but above the supply curve.

A. You can use integral calculus

B. If supply or demand are linear you can get it with geometry (and you can often approximate with geometry even if it isn't).

C. Our supply and demand charts actually show inverse demand and supply so it might be conceptually easier to use these when calculating surplus.

Now for one of the most celebrated results in economics:

The Invisible Hand Theorem: Absent externalities, CE maximizes efficiency.

A. That's right: if producers and consumers compete on price and achieve CE, then the market will:

   1. produce just the right amount of goods.
   2. give them to the people who value them most.
   3. even better, if anything changes, the CE outcome responds perfectly.

B. What's the qualification? **Externalities** refer to costs and benefits not captured or paid by the buyer and seller; we'll discuss examples later, like broadcast TV or water pollution.

C. Any other qualifications? Well, the market might not reach CE due to monopoly power, or info problems, or agency problems. Again, we'll discuss examples later.

D. Editorial remark: Some markets work quite well if left alone to self organize (e.g., rice). Others work well if engineered properly (e.g., electricity). A few seem problematic (e.g., police protection). What about personal data?

Ex: Using integration to find surpluses in with linear demand and supply

Deadweight Loss

I. The theory of competitive markets reveals the strength of markets but also (maybe more importantly) the unintended and often unseen downsides of common policy choices.

A. Bad policy leads to **deadweight loss**: it prevents the market from generating all of the surplus possible.

II. Price Controls

A. Price blocked from rising or falling to competitive equilibrium,

B. resulting in underproduction.

Ex: Rent control

III. Taxes

A. A more complex (and harder to see) cause of deadweight loss

B. Taxes separate the price consumers pay from the price suppliers receive.

1. No tax: $p_d = p_s$
2. Quantity tax (tax per unit): $p_d = p_s + t$
3. Value tax (tax on percentage spent): $p_d = (1 + t)p_s$
4. The equilibrium price $p^*(t)$ is different from the efficient price $p^*(0)$.

Ex: A quantity tax. Set $D = q^d, S = q^s$. Derive tax incidence formula

$$p_s(t) = p^* - \frac{t|D'|}{S' + |D'|}, \quad p_d(t) = p^* + \frac{tS'}{S' + |D'|}. \tag{4}$$

Also graph PS, CS, $T = tq^*(t)$, and DWL.

IV. Tariffs

A. A tariff is just a tax on a subset of sellers.

B. Although tariffs can be effective in aiding the untaxed subset, they cause an overall deadweight loss.

V. Subsidies

A. Even more counterintuitive, **subsidies** cause the same problem in reverse.

B. A subsidy to consumers for buying some product causes $p_d = p_s - t$

Elasticity

I. Demand and supply curves are often best described in terms of **elasticity**,

A. the proportional sensitivity to price.

B. Elasticity of Demand: $\epsilon_d = \frac{\partial \ln D}{\partial \ln p} = \frac{\partial D}{\partial p}\frac{p}{D}$

C. Elasticity of Supply: $\epsilon_s = \frac{\partial \ln S}{\partial \ln p} = \frac{\partial S}{\partial p}\frac{p}{S}$.

II. Is it Elastic?

A. If $|\epsilon| > 1$ we say the supply or demand curve is elastic.

B. If $|\epsilon| < 1$ we say the supply or demand curve is inelastic.

C. If $|\epsilon| = 0$ we say that supply or demand is perfectly inelastic

D. If $|\epsilon| = \infty$ we say that supply or demand is perfectly elastic

Ex: Perfectly elastic and inelastic curves.

III. Elasticity and Curves

A. Note that elasticity varies along a linear demand curves.

B. However log-linear supply and demand curves (described earlier) have constant elasticity (they are often called constant elasticity curves)

IV. Taxes and Elasticity

A. When a tax gets levied on a producer, how much of that tax ends up being "paid" by the consumer?

B. Turns out it depends on the elasticity of supply.

1. Extreme cases:
    - Perfectly inelastic: None of it gets passed on to consumers
    - Perfectly elastic: All of it gets passed on to consumers
2. In non-extreme cases it depends on the relative elasticity of the supply and demand curves.
3. In constant elasticity of demand markets you can find the effect of a tax on the price faced by consumers by applying a simple formula:

    $\frac{\partial p_d}{\partial t} = \frac{\epsilon_s}{|\epsilon_d| + \epsilon_s}$

    - The greater the elasticity of supply relative to the elasticity of demand, the greater the portion of taxes passed onto consumers.

The Communicative Role of Prices (Hayek, ...)

2. Technology and Cost

Based on Varian, Chapters 1, 4-6

I. Describing the Firm

A. The neoclassical specification of the firm is really just a description of the firm's production possibilities.

1. Which outputs can be obtained from given inputs
2. How much has to be spent to get those inputs
3. How these production possibilities generate cost curves for the firm.
4. These cost curves themselves completely describe *everything we need to know about the firm*, if we are neoclassical.
5. Obviously incomplete, but very useful...even to evolutionary biologists!

B. Input/Output

1. The firm produces a vector $\mathbf{y}$ of product quantities.
   a. We'll usually focus on a firm with a single product with quantity y.
2. The firm has a set of inputs it can use to create these products. We describe these inputs as a vector (or a bundle) $\mathbf{x} = (x_1, x_2, ..., x_n)$. Each x_i is the quantity of input i used for producing the output.

C. Technology

1. How much output y can the firm produce with input bundle $\mathbf{x}$?
2. The firm's *technology* specifies this:
3. The **input requirement set** V(y) consists of all of the bundles $\mathbf{x}$ that can produce output quantity y.

   Ex: Activity analysis and production plans.
   Basically, recipes. For spaghetti sauce, for 16Gb memory chips, for 100 rides to SFO, ...

4. The **production function** $y = f(\mathbf{x})$ describes the maximum output that can be produced with any input bundle .

   Ex: Cobb-Douglas technology, $y = a_o x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$.
   Ex: Leontief technology , $y = \min\{a_1 x_1, a_2 x_2, ..., a_n x_n\}$.

5. The **isoquant** for given output level y^* is the set of input bundles that can produce y^* . It is the lower boundary of $V(y^*)$.

D. Standard assumptions about technology

1. Monotone
   a. More input enables at least as much output.
   b. Say this using V's: if you can produce y' with $\mathbf{x}$ you can still produce y' with a bigger bundle $\mathbf{x}' \geq \mathbf{x}$.
      (The notation means $x_i' \geq x_i$ for each $i = 1, ..., n$.)
   c. This is innocuous if extra inputs can be thrown away, "free disposal."
2. Convex
   a. If plans $\mathbf{x}$ and $\mathbf{x}'$ are in $V(y)$ (i.e. can produce y), then so is the mixture $\alpha\mathbf{x} + (1-\alpha)\mathbf{x}'$, for any mixing proportion $0 < \alpha < 1$.
   b. If a production plan can be replicated, then it is reasonable to say that the technology is convex.

      Ex: Replicating two production plans to create convex hull $V(y)$.

3. Non-empty
   a. With enough of the right kinds of inputs, you can create any level of output y.
4. Closed: a boring technical condition.

E. Trade Offs in Production Plans

1. Assume we have a "smooth" production technology $y = f(x_1, ..., x_n)$.

2. At what rate can we substitute one of our inputs for another in producing a particular output, y?
   a. Called the **technical rate of substitution**.
   b. With a smooth technology with two inputs, it is just the slope of the isoquant.

   Ex: Using the production function (and taking total derivative), write the technical rate of substitution in terms of marginal products ($mp_i = \frac{\partial f}{\partial x_i}$):

$$TRS_{ij} = \frac{dx_j}{dx_i}|_{[f(\cdot)=y^*]} = -\frac{\partial f}{\partial x_i}/\frac{\partial f}{\partial x_j} = -\frac{mp_i}{mp_j} \tag{1}$$

   Ex: A Cobb-Douglas example.

3. **Elasticity of substitution** σ is elasticity of $[x_j/x_i]$ wrt $|TRS|$. It is a measure of isoquant curvature. See Varian for ugly details (optional).

F. Returns to Scale

1. The returns to scale tells us what happens when we try to scale up a production plan.
2. If we multiply $\mathbf{x}$ by $t > 0$, what happens to y?
3. Three cases:
   a. Constant returns to scale.
      - If $y = f(x_1, x_2)$, then for $f(tx_1, tx_2) = ty$
      - Output is proportional to the inputs.
   b. Increasing returns to scale.
      - If $y = f(x_1, x_2)$, then $f(tx_1, tx_2) > ty$ for $t > 1$.
      - We get more bang for our buck (at fixed prices of course) at higher scales of production.
   c. Decreasing returns to scale.
      - If $y = f(x_1, x_2)$, then $f(tx_1, tx_2) < ty$ for $t > 1$.

- We get diminishing returns from scaling our plans up.
- A major reason for DRS: there is some fixed input (not in the list), such as CEO attention, or planetary resources, or ...

Ex: Cobb-Douglas and returns to scale.

4. Homogeneity degree 1 as CRS. Homogeneous functions of degree $d = 0, 1, ...$

II. Cost Minimization

A. Behavior of the firm

1. We assume that firms economize in production:
2. they choose input bundles that minimize the cost of producing their chosen level of output.
3. When is this reasonable to assume? For competitive firms, and even for unrestrained monopolists. Only two exceptions come to mind:
   a. rogue managers pursue self-interest at the expense of firm owners, e.g., buy unnecessary corporate jets;
   b. old-fashioned regulators set price of a monopoly firm based on actual costs. Then it might be in the firm's interest to inflate costs.

B. The firm's problem.

1. To derive cost function, take as given the desired output quantity y, and the input price vector $\mathbf{w} = (w_1, w_2, ..., w_n)$. Sometimes also called factor prices.
2. Firms choose an input bundle $\mathbf{x}$.
   a. For convenience we will often write $\mathbf{x} = (x_1, x_2)$ and $\mathbf{w} = (w_1, w_2)$, but the reasoning extends to any finite vector of inputs.
3. The firm's main constraint (aside from factor prices) is technological.
   a. Can be summarized with the production function: $y = f(x_1, x_2)$.

4. So the firm's problem is simply:

$$c(\mathbf{w}, y) = \min_{x_1, x_2 \geq 0} w_1 x_1 + w_2 x_2 \text{ s.t. } f(x_1, x_2) = y \tag{2}$$

5. The four conditions noted earlier, including a strict version of convexity, allow us to say that if the problem above has an interior solution x^*, then it is unique and is characterized by the first order conditions, $w_i = \lambda mp_i, \quad i = 1, ..., n$.
6. Taking ratios of these first order conditions show that the technical rate of substitution is equal to the ratio of the factor prices: $|TRS_{ij}(x^*)| = \frac{w_i}{w_j}$.
7. The intuition is appealing: at the optimum input vector x^*, the isocost curve has the same slope (i.e., market tradeoff rate given by the price ratio) as the isoquant (the production tradeoff rate, TRS).
8. Even better, the Lagrange multiplier λ is equal to marginal cost! This can be seen from the general interpretation the shadow price, here of output in terms of expenditure on inputs. It can also be seen by solving any of the FOCs for λ, since w_i/mp_i is the cost of increasing output by a (micro) unit via increasing the input i. The insight (to be elaborated later) is that that cost must be the same for all inputs used in positive quantities.

C. Conditional factor demand

1. The firm's cost minimizing problem yields the firm's demand for each input as a function of prices and the scale of output.
2. Conditional factor demand for input i is $x_i^*(w_1, w_2, y)$.

D. The cost function

1. Cost functions represent the lowest cost of production available to a firm at a given set of factor prices. So we can rewrite equation (2) as

$$c(\mathbf{w}, y) = \mathbf{w} \cdot \mathbf{x}^*(\mathbf{w}, y) \tag{3}$$

2. With only two factors: $c(\mathbf{w}, y) = w_1 x_1^*(w_1, w_2, y) + w_2 x_2^*(w_1, w_2, y)$

Ex: Constant Elasticity technology.

Special cases: Cobb-Douglas technology, Leontief technology, Linear technology

E. Relationship between cost and conditional factor demand

1. If the cost function is differentiable, then you can use it to recover the input (or factor) demand functions.
2. This is known as **Shephard's lemma**: $x_i^*(\mathbf{w}, y) = \frac{\partial c(\mathbf{w}, y))}{\partial w_i}$.
3. To verify, differentiate equation (3) wrt w_i. The main effect is as in Shepard's lemma, but there is also an indirect effect $\mathbf{w} \cdot \frac{\partial x_i^*(\mathbf{w}, y))}{\partial w_i}$. By the FOC, this indirect effect is proportional to $(mp_1, ...mp_n) \cdot \frac{\partial x_i^*(\mathbf{w}, y))}{\partial w_i} = 0$, as can be seen by differentiating the isoquant identity $f(\mathbf{x}^*(\mathbf{w}, y)) = y$.
4. This vanishing indirect effect is an example of the envelope theorem. Geometrically, the idea is that the production function gradient (i.e., marginal product vector) is normal to (aka orthogonal or perpendicular to) the isoquant surface, and therefore also normal to the isocost surface, so the indirect effect is zero. See Varian p. 74 for further discussion.

F. Duality and properties of cost functions

1. Suppose that you have a function $c(\mathbf{w}, y)$ with the following properties
   a. monotone increasing in each argument,
   b. homogeneous degree 1 in $\mathbf{w}$,
   c. concave in $\mathbf{w}$, and
   d. continuous and (at least piecewise) differentiable

   Then there is some nice [monotone, ..., closed] production function for which c is the cost function!

   Conversely, if c is the cost function for some nice production function, then it satisfies the four properties just mentioned.

2. Implications for applied work:

   (a) usually you can skip estimating a production function, especially if not all data on input quantities are available, and just estimate the cost function directly, using data on input prices and output levels, which usually is easier to collect.

   (b) when estimating a cost function, consider imposing homogeneity and monotonicity as coefficient restrictions, and testing for concavity.

Ex: CES cost function — see Appendix.

Ex: Translog cost function

$$\ln c = a_0 + a_1 \ln w_1 + a_2 \ln w_2 + \frac{1}{2} b_{11} [\ln w_1]^2 + b_{12} \ln w_1 \ln w_2 + \frac{1}{2} b_{22} [\ln w_2]^2 \ , \quad (4)$$

where homogeneity of degree 1 implies (a) $a_1 + a_2 = 1$ and (b) $b_{11} + b_{22} + 2b_{12} = 0$. See Varian pp. 210 for factor share calculations.

III. Cost Curves with a single variable input

A. Cost curves

1. Focusing from here on cost instead of production, let's consider short vs long run.
2. The main ideas come through most clearly when we assume just two inputs to production which we will rename f (for fixed in SR) and v (for freely variable).
   a. $\mathbf{x} = (x_f, x_v)$
   b. $\mathbf{w} = (w_f, w_v)$

B. The short run cost curve

1. In the short run, some of the costs are fixed.
   a. In the short run we can't buy new machinery, build new factories, hire new management, or change union contracts.

b. Those costs are fixed – we represent them as F.

c. Part of F can be recovered if we halt production (e.g.s). This part is called **avoidable**. the remainder is called **sunk**.

d. Buried here is a general point: economic costs = opportunity costs, not necessarily cash or accrual costs. Ask yourself: does F change in the SR when w_f increases?

2. Total costs incurred by the firm consist of both variable costs and fixed costs (which are in turn simple to express in the two inputs model).

   a. Variable Cost: c_v

      - $c_v(y) = w_v x_v(w_v, y, x_f)$

      Ex: Variable Cost Curve

   b. Fixed Cost: F

      - $F = w_f x_f$

      Ex: Fixed cost curve

   c. Total Cost: $c_v(y) + F$

      - $c(y) = w_v x_v(w, y, x_f) + w_f x_f$

      Ex: Total cost curve

3. Average Costs

   a. There are three main types of average costs that are used in studying firm behavior.

      - Average Total Cost: $AC(y) = \frac{c(y)}{y} = \frac{c_v(y)+F}{y}$
        - U-shaped
      - Average Variable Cost: $AVC(y) = \frac{c_v(y)}{y}$
        - Eventually rising

- Average Fixed Cost
  - Always falling

b. The three average cost functions are related by a simple equation:

c. $AC(y) = AFC(y) + AVC(y)$

Ex: Average cost curves.

4. Marginal Costs
   a. $MC(y) = \frac{\partial c}{\partial y} = \frac{\partial c_v}{\partial y}$
   b. Relationship between MC and AC.
      - When AC is decreasing, MC must be smaller than AC.
      - When AC is increasing, MC must be larger than AC.
      - MC intersects AC at the minimum point of the AC.
        - Minimum efficient scale
      - The same must be true with AVC!
      - Integral of (area under) MC gives VC.

Ex: Suppose TC is $c(y, \mathbf{w}) = 128 + 69y - 14y^2 + y^3$ for some fixed input price vector $\mathbf{w}$. Find FC, MC, VC, AVC, etc. Where appropriate, assume all fixed costs are sunk.

To look ahead a bit, find the short run supply curve by solving $p = MC_+(y)$ for y. That is, obtain $y^*(p, \mathbf{w})$ from the increasing portion MC_+ of the marginal cost curve. (Later we will see that not all of MC_+ is relevant, just the part above the AVC curve.)

C. Long run cost curve

1. As we saw in the last section, decreasing returns set in when a factor is fixed.
2. Eventually, the cost of short run factors force the firm to change long run factors.
3. Relationship between long and short term curves

a. Suppose costs are given as a function of output and the corresponding demand for the fixed factor, $x_f(y)$: $c(y, x_f(y))$.

b. Fix y at some level $\bar{y}$ and refer to $\bar{x}_f$ as the optimal amount of the fixed factor for producing $\bar{y}$.

c. Now lets differentiate $c(\bar{y}, x_f(\bar{y}))$ with respect to y at $\bar{y}$.

- This gives us the following:

$$\frac{dc(\bar{y},x_f(\bar{y}))}{dy} = \frac{\partial c(\bar{y},\bar{x}_f)}{\partial y} + \frac{\partial c(\bar{y},\bar{x}_f)}{\partial x_f}\frac{\partial x_f(\bar{y})}{\partial y}$$

d. **Envelope theorem**

- The conditional factor demand, x_f is the cost minimizing factor choice for producing $\bar{y}$.
- Thus $\frac{\partial c(\bar{y},\bar{x}_f)}{\partial x_f} = 0$
- Plugging this into the math above we get a great expression for the relationship between long run cost and short run cost.

$$\frac{dc(\bar{y},x_f(\bar{y}))}{dy} = \frac{\partial c(\bar{y},\bar{x}_f)}{\partial y}$$

- That is the slope of long run cost equals the slope of short run cost! (at the point of tangency)
- The long term costs are the lower envelope of all of the short run costs at various levels of production.

D. Learning curve

1. Experience may enable a firm to discover better procedures and techniques, avoid waste, etc.
2. First quantified in WWII aircraft and shipbuilding, true in teaching classes, manufacturing memory chips, etc etc.
3. The usual specification is in accumulated output $Y_t = \sum_{s \leq t} y_s$ that AC falls proportionately, $\ln AC_t = AC_0 - b \ln Y_t$.

E. Multiproduct firms.

1. Suppose that the joint cost function (estimated directly) is $c(y_1, y_2; \mathbf{w})$.
2. **Economies of scope** exist if $c(y_1, y_2; \mathbf{w}) < c(y_1, 0; \mathbf{w}) + c(0, y_2; \mathbf{w})$, e.g., because distribution networks can be shared, or R&D, or production facilities.
3. Another reason is the presence of cost complementarities, $\frac{\partial^2 c}{\partial y_1 \partial y_2} < 0$, i.e., increasing the output of one product lowers the MC of the other output. E.g., Big Creek Lumber product 1=redwood siding, product 2 = redwood sawdust.

Ex: Multiple plants – a lead-in to supply curve aggregation.

IV. Appendix: The CES cost function

Consider the cost function

$$c(y, w_1, w_2) = [(\frac{w_1}{a_1})^r + (\frac{w_2}{a_2})^r]^{\frac{1}{r}} y, \tag{5}$$

where y is the input quantity and the w_i's are the input prices.

- One special case is $r = 1$. This gives linear iso-cost curves; the inputs are perfect substitutes.
- Another is $r = 0$. Use L'Hospital's rule etc to see that this gives Cobb-Douglas iso-cost curves
- For $0 < r < 1$, the iso-cost curves intersect the axes; the inputs are imperfect substitutes but neither is essential.
- For $-\infty < r < 0$, the iso-cost curves don't intersect the axes; both inputs are essential.
- As $r \to -\infty$, we get Leontieff iso-cost curves; inputs needed in fixed proportions.

The corresponding production function is

$$f(x_1, x_2) = [(a_1 x_1)^\rho + (a_2 x_2)^\rho]^{\frac{1}{\rho}}, \tag{6}$$

where r and ρ are "dual": $\frac{1}{r} + \frac{1}{\rho} = 1$. See Varian p. 55-56. In particular,

- $r = 1 \iff \rho = -\infty$,
- $r \downarrow 0 \iff \rho \uparrow 0$
- $r = -\infty \iff \rho = 1$

so [linear, C-D, Leontieff] cost corresponds to [Leontieff, C-D, linear] production!

CES stands for constant elasticity of substitution. That elasticity, denoted σ, is a measure of iso-cost curvature; see Varian p. 20.

To estimate σ, look at the first order conditions for cost minimization and do some algebraic manipulations to obtain

$$\ln \frac{x_1}{x_2} = a_0 + \sigma \ln \frac{w_1}{w_2}, \quad \text{where} \quad a_0 = -\frac{\sigma}{\rho} \ln \frac{a_2}{a_1}. \tag{7}$$

The discussion so far assumes constant returns to scale. More generally, replace the outer exponent $\frac{1}{\rho}$ in equation (6) by $\frac{\alpha}{\rho}$. Of course, $\alpha > 1$ (or < 1) specifies increasing (or decreasing) returns to scale.

Extra credit for the mathematically ambitious: find the cost function and estimating functions corresponding to $\alpha \neq 1$.

3. Profit and Supply.

See Varian Ch 2-3

I. Maximizing Profit

A. Neoclassical theory assumes that all firms act so as to maximize profits.

B. As noted in Varian's nice discussion at the beginning of Ch 2, this immediately implies that the marginal revenue for any activity the firm undertakes must equal its marginal cost. This insight alone is enough to guide some decisions (or to solve certain homework problems)!

C. You will recall that *competitive* firms are (by definition) price takers. This applies to output(s) as well as inputs.

D. Maximizing profit implies that the firm minimizes cost: whatever output it might choose, the firm can only maximize profit if it minimizes the cost of producing that output. This is true (as we will discuss later) even for firms that are not competitive on the output market.

E. Here's some math to support that intuitive statement, and shed additional light on SR profit maximization.

1. For clarity, again we hold one factor f fixed at some level $\bar{x}_f$, and let the firm freely choose the level $x_v \geq 0$ of another factor v.

2. Use the symbols π, R, p to represent profit, revenue, output price, and keep y as output level and w's as input prices. Since profit = revenue - cost, the firm's problem is

$$\begin{aligned} \max_{y,x_v \geq 0} \pi &= \max_{y \geq 0}[\max_{x_v \geq 0} R(y) - w_v x_v - w_f \bar{x}_f \text{ s.t. } y = f(x_v, \bar{x}_f)] \qquad (1) \\ &= \max_{y \geq 0}[R(y) - \min_{x_v \geq 0}[w_v x_v - w_f \bar{x}_f \text{ s.t. } y = f(x_v, \bar{x}_f)]] \\ &= \max_{y \geq 0}[R(y) - c(y)] \end{aligned}$$

where $c(y)$ is shorthand for the SR cost function (which also depends on the w's and on $\bar{x}_f$).

3. If the firm is competitive (a price-taker), then $R(y) = py = pf(x_v, \bar{x}_f)$.

4. To analyze *un*conditional factor demand, write the competitive firm's problem as

$$\max_{x_v \geq 0} pf(x_v, \bar{x}_f) - w_v x_v - w_f \bar{x}_f. \tag{2}$$

5. The FOC is $p\frac{\partial f}{\partial x_v} = w_v$, i.e., the value of the marginal product must be equal to its price!

Ex: Graphical representation of problem (2).

Ex: Factor demand with Cobb-Douglas production function.

F. Price changes

1. Increasing the price of a factor will obviously decrease the firm's factor demand.

   Ex: Differentiating FOC by w.

   Ex: Graphical representation of factor price increase.

2. Increasing the price of output y however will increase the firm's factor demand.

   Ex: Differentiating FOC by p.

   Ex: Graphical representation of product price increase.

G. The long run version

1. In the long run the problem is exactly the same except that the firm chooses both factors.

$$\max_{x_v, x_f \geq 0} pf(x_v, x_f) - w_v x_v - w_f x_v \tag{3}$$

2. Values of each marginal product have to equal respective prices.

II. Profit functions, supply and (factor) demand

A. We can define a *profit function* as the solution to problem (3), i.e.,

$$\pi(p, \mathbf{w}) = \max_{\mathbf{w} \geq 0} pf(\mathbf{x}) - \mathbf{w} \cdot \mathbf{x}. \tag{4}$$

B. Warning: this profit function does not always exist – corners and kinks complicate its expression, and IRS or even CRS production functions cause problems. Cost functions exist under much milder assumptions than profit functions.

C. In applied work, one sometimes sees profit functions estimated directly from data. (E.g., in banking, and for internet platforms like iTunes or even Facebook.) If so, it is helpful to know the general properties such a function must have, so they can be imposed as parameter restrictions in the estimation.

D. Using straightforward arguments, Varian shows that profit functions (when they exist) must be (a) (weakly) increasing in p and decreasing in each w_i, (b) homogeneous of degree 1 in $(p, \mathbf{w})$, (c) convex in p, and (d) continuous in p.

E. One great thing about cost functions is that they tell us conditional factor demands (via Shephard). Profit functions, when they exist, can tell us unconditional factor demands, and even the supply function.

F. Hotelling's Lemma tells us that the supply function is:

$$y^*(p, \mathbf{w}) = \frac{\partial \pi(p, \mathbf{w})}{\partial p} \tag{5}$$

and that unconditional factor demands are

$$x_i(p, \mathbf{w}) = -\frac{\partial \pi(p, \mathbf{w})}{\partial w_i} \tag{6}$$

G. The proof is analogous to that for Shephard's lemma, and involves the envelope theorem; see Varian p43-44 for self-contained proofs.

Ex: The one input, one output case.

III. The Firm's Supply

A. The firm's supply curve is $y(p) = y^*(p, \mathbf{w})$ holding constant the input price vector $\mathbf{w}$ while letting output price vary.

B. The supply curve turns out to be just selected portions of the marginal cost curve. To see this notice that

   1. The competitive firm's profit maximization problem (taking output price and factor prices as given) can be expressed as
      $\max \pi = py - c(y)$
   2. The FOC is
      $p = c'(y)$
   3. The firm sets marginal cost equal to marginal revenue as usual, but MR = p for a competitive firm!
   4. This is just the inverse supply curve that we regularly graph when doing competitive analysis.

C. Some qualifications

   1. "U" shaped marginal cost curves
      a. If MC is U-shaped, then there may be two levels of output where price and marginal cost are equal.
         - In fact when MC is decreasing, the gap between revenue and cost is increasing as more is produced.
         - If this is so, the firm will always produce more.
         - This is confirmed by the SOC $0 < \pi''(y) = -c''(y)$, associated with a relative minimum.
      b. It is only upward sloping portions of MC (where $c''(y) > 0$ and thus the SOC $\pi''(y) < 0$ holds) that can be part of the inverse supply curve.
   2. The Shutdown Condition
      a. Firms incur (sunk) fixed costs even if they aren't producing anything, and this fact can encourage them to produce even when they can't turn

a positive profit. Here's the algebra, where y denotes some positive level of output:

$$\begin{aligned} \pi(0) &> \pi(y) \qquad \Longleftrightarrow \qquad (7) \\ -F &> py - c_v(y) - F \iff \\ AVC = \frac{c_v(y)}{y} &> p \end{aligned}$$

b. This implies that the firm is only willing to produce if the price is above average variable cost.

3. Conclusion: the firm's supply function $y^*(p)$ follows the increasing portion of the MC curve above AVC; elsewhere $y^*(p) = 0$.

D. An Individual Firm's Producer Surplus

1. Recall that producer surplus is the area above the supply curve, below the price line.
   - The same is true of the individual firm.
   - Recall that the area under the marginal cost curve is the variable cost curve.
   - Thus the producer surplus is the difference between total revenue and total variable cost!

   Ex: Graphical representation of PS with the individual firm.

2. It shouldn't be surprising that profits have a natural relationship to producer surplus.
   a. Profits are just the difference between total revenue and total cost

   $\pi = py - c_v(y) - F$

   b. Surplus is the difference between total revenue and variable cost

   $PS = py - c_v(y)$

c. Thus producer surplus is just the sum of profits and fixed costs.

$PS = \pi + F$

IV. Competitive Industry

A. Short run supply

1. The industry supply curve is just the **horizontal** sum of individual market supply curves.
   a. Sum of quantities supplied.
   b. To the individual firms, prices are exogenous.
   c. Prices individually determine firms' output decisions.
2. In the short run,
   a. Sunk fixed costs F are incurred no matter what.
   b. As long as price is above its AVC, each firm is better off staying in business even if taking losses.
   c. No entry or exit on this time scale (starting a business usually involves incurring fixed costs).
3. Use the same reasoning for multi-plant firms.
   a. For two plants, $\max R(y_A + y_B) - c(y_A) - c(y_B)$, and look at FOCs in competitive case $R(y_A + y_B) = (y_A + y_B)p$.
   b. In SR, keep each plant operating at the output level where $p = MC(y_i)$, on increasing branch of MC curve above AVC.
   c. Shut down right away any plant that can't cover its VC.
      (In LR, shut down or sell off any plant that can't cover its total cost.)
   d. replace p by MR if the firm has market power.

B. Long run supply

1. In the long run, firms can adjust all of their productive inputs.

a. Firms can enter

b. Firms can exit without incurring fixed costs.

- Now the supply curve is marginal cost in excess of *average cost*.

2. Entry

a. When there are few firms in a market, they each may make relatively high profits.

- Profits attract new firms.
- Each new (identical) firm in a market causes the supply curve to become more elastic.
- This in turn lowers prices and therefore the profits of individual firms.

b. This process of attraction and entry only ends when there are so many firms in the market that an extra entrant would drive profits below zero.

- Meaning that, in the long run, price will be driven close to minimum average cost.
- The LR supply curve is thus infinitely elastic at minimum average cost.

c. The long run cost firm of a competitive industry is flat, exhibiting constant returns to scale.

d. Profits are signals that mobilize entrepreneurs to focus attention on highly valued projects.

e. In LR eq of competitive industry, the econ profit is zero, meaning that accounting profit (ROE) is at normal level (WACC).

C. Rent and rent seeking

1. Rent is the payment to a factor in excess of its opportunity cost.

a. Determined by the market price for outputs and variable costs of production.

b. Rent is really just producer surplus seen in a different light.

2. A competitive market for the factor will drive the price to the full amount of the rent.
    a. Drives firms' profits back down to zero.
3. Politics is a second way rent can be dissipated.
    a. Firms should be willing to spend all of the potential rents to control fixed factors.
    b. By creating or maintaining barriers to entry firms can create artificial scarcity.
    c. The rents from this scarcity are valuable and firms will pay to create it.
    d. This is pure waste and is called *rent seeking*.

4. Preferences and Demand

See Varian Chapters 7-9

I. Preference Relations and Utility Functions

A. Our main goal in this section is to model how people choose among consumption opportunities.

B. Each opportunity is called a **bundle**.

C. A bundle is represented as a vector $\mathbf{x} \in \mathbf{R}^n_+$ with dimension n equal to the number of goods. On Amazon, $n > 100,000$.

D. But the main ideas can be conveyed with $n = 2$, where $\mathbf{x} = (x_1, x_2)$.

- $\mathbf{x} = (7, 4)$ is the bundle consisting of 7 units of good 1 and 4 units of good 2.
- E.g., $x_1 = \#$ of cowboy wraps at lunch truck, and $x_2 = \#$ sodas.
- Or, $x_1 =$ number of useful ideas in the lecture, and $x_2 =$ style points.

Ex: Bundles in a 2-dimensional consumption space $\mathbf{R}^2_+$.

E. Preferences Over Bundles

1. We assume that people can compare alternative bundles, and know which they like better.
2. A **preference** is a relation between between pairs of bundles.
3. Taking a consumer's response to any two bundles $\mathbf{x}$ and $\mathbf{y}$, we write
   - $\mathbf{x} \sim \mathbf{y}$ ("indifference") if she is just as happy with $\mathbf{x}$ as she is with $\mathbf{y}$.

   OR

- $\mathbf{x} \succ \mathbf{y}$ or $\mathbf{x} \prec \mathbf{y}$ ("strict preference") if she either happier with $\mathbf{x}$ than $\mathbf{y}$ or is happier with $\mathbf{y}$ than $\mathbf{x}$.
- $\mathbf{x} \succeq \mathbf{y}$ (or $\mathbf{x} \preceq \mathbf{y}$) means that this consumer is either indifferent between $\mathbf{x}$ and $\mathbf{y}$ or strictly prefers $\mathbf{x}$ to $\mathbf{y}$ (or she is either indifferent between y and $\mathbf{x}$ or strictly prefers $\mathbf{y}$ to $\mathbf{x}$). This is called **weak preference** or just plain **preference**.

Ex: Indifference curves (IC's) and better sets in a 2-dimensional consumption space.

4. Preferences are defined by three properties:
   a. Complete: Either $\mathbf{x} \succeq \mathbf{y}$, or $\mathbf{y} \succeq \mathbf{x}$, or both (in which case $\mathbf{x} \sim \mathbf{y}$).
   b. Reflexive: $\mathbf{x} \succeq \mathbf{x}$.
   c. Transitive: If $\mathbf{x} \succeq \mathbf{y}$ and $\mathbf{y} \succeq \mathbf{z}$ then $\mathbf{x} \succeq \mathbf{z}$.

   Ex: Preference no-no: Indifference curves that cross.

5. Preferences are often assumed to have additional properties. Three of the most important are:
   a. Monotone (increasing): More is better: $x_i \geq y_i \ \ \forall i \implies \mathbf{x} \succeq \mathbf{y}$.
   b. Convex: If $\mathbf{x} \sim \mathbf{y}$ then for any $0 \leq \alpha \leq 1$,
      $(\alpha x_1 + (1 - \alpha y_1), \alpha x_2 + (1 - \alpha y_2)) \succeq (x_1, x_2)$ (Mixtures are better)
   c. Inada: $[x_i > 0 \ \ \forall i]$ & $[y_i = 0$ for some $i] \implies \mathbf{x} \succ \mathbf{y}$.

   Ex: Non-monotone prefs. Non-convex prefs. Non-Inanda prefs.

F. Utility Functions.

The preference theory we just talked about seems cumbersome — you wouldn't want to list all possible pairs, their relation, and check whether the properties

hold ! So it is very nice that we have utility functions to work with.

- If preferences (which are automatically complete, reflexive and transitive) are also continuous (a property awkward to define formally, but intuitively clear) and monotone, then it is known that we can represent those preferences by a continuous utility function. [Varian, p.97 sketches a proof.]
- A **utility function** u assigns a real number $u(\mathbf{x})$ to each bundle $\mathbf{x}$.
- The utility function u **represents** preferences $\succeq$ if
  1. $\mathbf{x} \succ \mathbf{y}$ if and only if $u(\mathbf{x}) > u(\mathbf{y})$, and
  2. $\mathbf{x} \sim \mathbf{y}$ if and only if $u(\mathbf{x}) = u(\mathbf{y})$.
- So a utility function is just like a production function whose inputs are bundles and whose output is the degree of satisfaction. From a mathematical perspective, the only difference is that
- the *same* preferences can be represented by several *different* utility functions.
  1. Any monotone increasing transformation v of a utility function u represents the same preferences as u, i.e. you can use either one, whichever is more convenient.
  2. Basically, by applying v, you leave the IC's the same and their ordering the same, but just change the labels on the IC's from $u =$ const to $v =$ some other constant. (Don't do this with production functions !!)
  3. No harm in choosing a smooth (continuously differentiable) utility function.
- The partial derivative of a utility function is its **marginal utility**: $mu_i \equiv \frac{\partial u}{\partial x_i}$. It depends on the choice of a utility function, but, more importantly...

- The **marginal rate of substitution** between two goods i and j is $MRS_{ij} \equiv -mu_j/mu_i \equiv -\frac{\partial u}{\partial x_j}/\frac{\partial u}{\partial x_i}$.

    1. MRS_{ij} is the consumer's trade-off rate between those two goods. It is how many (micro)units of i that the consumer requires to just compensate for losing one (micro)unit of j.
    2. MRS_{ij} is invariant to the choice of utility function to represent given preferences!
    3. The Inada property mentioned earlier implies that

       $MRS_{ij}(\mathbf{x}) \to 0$ (or ∞) as x_i (or x_j) $\to 0$.

       The idea is that IC's don't intersect axes for goods that are essential.
    4. MRS_{21} is just the slope of the indifference curve in (x_1, x_2) space.
    5. In higher dimensions, there is an indifference hypersurface, and there MRS_{ij} is the slope of that surface in the $j - i$ plane.

Ex: Perfect substitutes: $u(x_1, x_2) = x_1 + cx_2$. Then $MRS_{12}(x_1, x_2) = c > 0$; the tradeoff rate is constant.

Ex: Cobb-Douglas utility: $u(x_1, x_2) = \ln x_1 + c \ln x_2$. Then $MRS_{12}(x_1, x_2) = \frac{cx_1}{x_2} > 0$. Convex, Inada. What can we say about $v(x_1, x_2) = \exp(u(x_1, x_2))$?

Ex: CES utility: $u(x_1, x_2) = \frac{1}{\rho}\ln(x_1^\rho + cx_2^\rho)$, where $\rho \in (-\infty, 1]$. Can show that this nests the previous two cases, which correspond respectively to $\rho = 1, 0$. Generalizes directly to more than 2 goods. Very useful in applied work.

Ex: Quasilinear utility: $U(x_0, x_1) = x_0 + g(x_1)$. Think of good 0 as money, or purchasing power. $MRS_{ij}(x_0, x_1) = g'(x_1)$. Very very useful in applied work.

II. The Direct Consumer Problem and (Marshallian) Demand

Now we can model choice and see where demand functions come from. Let me follow textbooks for the next hour. Later I will show you a streamlined approach I developed with Jozsef Sakovics (with roots in Marshall).

Suppose that the consumer is constrained only by available income and by the prices of the goods.

A. Let m be the available money to spend, while p_1 the price of good 1, p_2 the price of good 2 etc.

B. Then the **budget constraint** is

$$m \geq \mathbf{p} \cdot \mathbf{x} = p_1 x_1 + p_2 x_2 + ..., \tag{1}$$

(i.e. you can't spend more than you have.)

C. If she has strictly monotone preferences, a consumer will spend all of her money — putting money in a savings account could count as one of the x_i's, and m as such brings no utility. So we can safely assume that equation (1) holds as an equality:

$$m = \mathbf{p} \cdot \mathbf{x} = p_1 x_1 + p_2 x_2 + ... \tag{2}$$

Ex: Budget constraints and budget sets.

D. Then we have the following constrained optimization problem.

$$\max u(x_1, x_2, ...) \quad \text{s.t. } m = p_1 x_1 + p_2 x_2 + ... \tag{3}$$

We form the Lagrangian

$$\mathcal{L} = u(x_1, x_2, ...) + \lambda(m - p_1 x_1 + p_2 x_2 + ...) \tag{4}$$

Differentiating, we get first order conditions:

$$\begin{aligned} 0 &= \frac{\partial u}{\partial x_1} - \lambda p_1 \implies mu_1 = \lambda p_1 \\ 0 &= \frac{\partial u}{\partial x_2} - \lambda p_2 \implies mu_2 = \lambda p_2 \\ 0 &= ... \\ 0 &= \frac{\partial \mathcal{L}}{\partial \lambda} \implies m = p_1 x_1 + p_2 x_2 + ... \end{aligned} \tag{5}$$

E. Simultaneously solving these FOCs gives us optimal consumption decisions, at least if standard assumptions hold. (Namely, strong monotonicity, smooth indifference curves, Inada and convexity ensure that (5) has a unique solution that solves the consumer choice problem (3). Even without Inada and convexity one can often get an interior solution to (3) where (5) holds.).

F. Without actually solving the FOCs (5), dividing condition i by condition j gives us a familiar expression:

$$p_i/p_j = \frac{\partial u}{\partial x_i} / \frac{\partial u}{\partial x_j} = MRS_{ji}. \tag{6}$$

G. That is, the budget line and indifference curve have the same slope at the optimal bundle and, given the last condition in (5), they are actually tangent at that point.

Ex: An example with Cobb-Douglas preferences.

Ex: Direct consumer problem with indifference curves.

Ex: Corner solutions. Non-convex prefs. non-monotone prefs.

H. Solving problem (3), sometimes called the *direct consumer problem*, gives us the quantity x_i demanded for each good i as a function of prices and income. This is the individual's **Marshallian demand** $x_i^*(\mathbf{p}, m)$.

III. Comparative statics of Marshallian demand.

If we hold the price of other goods and income constant in $x_i^*(p_1, p_2, ..., m)$, we get (for that individual consumer) the old demand curve we studied in the first week $x_i(p_i)$. It shifts with changes in other prices and changes in income.

Ex: Marshallian demand from Cobb-Douglas preferences. Obtain

$$\ln x_i = A + e_{ii} \ln p_i + e_{ij} \ln p_j + \eta_i \ln m \tag{7}$$

with $e_{ii} = -1, e_{ij} = 0, \eta_i = 1$, and $A =$ ln expenditure share.

A. Income sensitivity: $\frac{\partial x_i^*}{\partial m}$

We just saw that m affects the quantity demanded. What happens when income (actually, expenditure) changes? Of course, it depends on the structure of preferences – we'll go through a few of the possible cases, drawing the Income Expansion Paths (IEPs or, better, EEPs).

- **Normal goods** are goods that consumers buy more of as income increases: $\frac{\partial x_i^*}{\partial m} > 0$.
- **Inferior goods** are goods that consumers buy less of as income increases: $\frac{\partial x_i^*}{\partial m} < 0$.

- With **quasilinear** preferences and sufficient income, consumers don't change the amount they buy as income increases: $\frac{\partial x_i^*}{\partial m} = 0$.

Ex: To explain that last point, suppose that $u(x_0, x_1) = x_0 + g(x_1)$. Check that all extra money gets spent on good 0 (i.e., kept as cash) and none on good 1, once enough x_1 has been purchased so that $g'(x_1) \leq p_1$.

B. If preferences are **homothetic**, then IEPs are straight lines, so the proportion of total consumption represented by each good remains the same regardless of income.

1. If m doubles, then consumption of each good doubles.
2. If you multiply m by t, then consumption of each good gets multiplied by t
3. unit income elasticity: $\frac{\partial \ln x_i^*}{\partial \ln m} = 1$.

Ex: Cobb-Douglas preferences are homothetic.

C. Price Changes: $\frac{\partial x_i^*}{\partial p_i}$.

Marshallian demand is more complicated than you might think when it comes to the basic question of own-price effects. We'll go through the messy (but standard) approach now, and later will see how other versions of demand make things more intuitive.

Two different effects are at work (sometimes in opposite directions):

- **Substitution effect**: A change in the price ratio affects consumers' trade-off rates among goods.
- **Income effect**: The consumer's spending power (her real income) changes because goods overall are cheaper (if the price went down) or more expensive (if the price went up).

1. Income Effect. As we just saw, the income effects of price changes can be either positive or negative depending on whether the price increase or decreases and whether the good is normal or inferior.
2. Substitution Effect. You can isolate the substitution effect by imagining what a consumer would choose if the price change happened and then we either gave the consumer some money (if the price increased) or took away some money (if the price decreased) to just compensate for the change in purchasing power.
3. There are several slightly different methods to separate the substitution effect from the income effect; we'll just do one, called the **Hicks decomposition**.
4. What if we add or subtract just enough income so that the consumer's optimal consumption after the price change, though different, is just as pleasing to him as before?
5. The change in consumption due to the change in price *after* so compensating her income is a measure of the substitution effect.

   Ex: Graphical example: roll and shift.

6. Hicksian Demand
   - There is an alternative version of the demand function called **Hicksian demand** and denoted $h_i(\mathbf{p}, u)$, that consists only of this compensated demand.
   - It is defined as the quantity of the good a consumer demands when choosing the cheapest (expenditure minimizing) bundle she could buy while maintaining a given level u of utility.

- Hicksian demand charts only changes in quantity demanded that are due to substitution effects.
- Hicksian demand is always inversely related to price, i.e., satisfies $\frac{\partial h_i(\mathbf{p},u)}{\partial p_i} < 0$.

7. **The Slutsky Equation** formally shows how to decompose $(\frac{\partial x_i^*}{\partial p_i})$ into an income and substitution effect. Below $e(\mathbf{p}, u)$ is the minimum expenditure m necessary to acquire a bundle that brings utility level u given price vector $\mathbf{p}$. It is an easy consequence of the envelope theorem that $\frac{\partial e}{\partial p_i} = x_i^*$.
   - Start with the identity $h_i(\mathbf{p}, u) = x_i^*(\mathbf{p}, e(\mathbf{p}, u))$.

     Differentiate both sides with respect to p_i and rearrange terms to get the Slutsky equation

     $$\frac{\partial x_i^*(\mathbf{p}, m)}{\partial p_i} = \frac{\partial h_i(\mathbf{p}, u)}{\partial p_i} - \frac{\partial x_i^*}{\partial m} x_i^*(\mathbf{p}, m). \tag{8}$$

   - There are two terms on the right hand side (RHS) of the Slutsky equation:
     a. The first term is the substitution effect. It is the derivative of Hicksian demand evaluated at $u =$ the maximum level of utility achievable given the prices and income.
     b. The second term $-\frac{\partial x_i}{\partial m} x_i(\mathbf{p}, m)$ is the income effect. It turns out to be the income sensitivity found in the previous section weighted by the quantity currently demanded.

8. **The "Law of Demand"** states that $\frac{\partial x_i^*(\mathbf{p},m)}{\partial p_i} < 0$, i.e., that demand curves slope downwards. It is true for all normal goods. Why?
   - The substitution effect is always negative as we have just seen, so the first term on the RHS of (8) is negative.

- If the good is normal, then the second term on the RHS of (8) is also negative, so the entire RHS is negative, as claimed.

9. Giffen Goods

   a. Usually substitution effects are greater than income effects and so, even if goods aren't normal, demand still slopes downward.

   b. However, it is theoretically possible for a good to be so inferior that they overcome the substitution effect and, in places demand slopes up.

   c. These theoretical constructs are called **Giffen goods**.

Ex: Graphical example of a Giffen good.

IV. Duality and other tools.

Pursuing the analogy to the supply side, we can construct functions dual to demand and utility. It is conceptually helpful, though perhaps not quite as important in empirical work.

A. **Indirect utility** is $v(\mathbf{p}, m) = [\max u(\mathbf{x}) \text{ s.t. } m = \mathbf{p} \cdot \mathbf{x}] = u(\mathbf{x}^*(\mathbf{p}, m))$.

B. Dual problem to utility max is expenditure min:

$$\min_{\mathbf{x}} \mathbf{p} \cdot \mathbf{x} \text{ s.t. } u(\mathbf{x}) \geq u_o. \tag{9}$$

C. Solution to (9) is called expenditure function, denoted $e(\mathbf{p}, u_o)$, and argmax's are called Hicksian demands, denoted $h_i(\mathbf{p}, u_o)$, used earlier.

Ex: indirect utility, expenditure and Hicksian demand from Cobb-Douglas preferences.

D. see Varian for a bunch of general properties and identities, some of which can help in applied work.

E.g., $h_i(\mathbf{p}, u_o) = x_i^*(\mathbf{p}, e(\mathbf{p}, u_o))$, interpreted as compensated demand, as we will see shortly.

E. **Elasticity form of the Slutsky equation**: Multiply both sides of equation (8) by $\frac{p_i}{x_i}$, also multiply the last term by $\frac{m}{m}$, and simplify, using the usual expressions for elasticities. Writing $s_i = \frac{p_i x_i}{m}$ as the expenditure share of good i we get

$$\epsilon_i = \epsilon_i^h - s_i \epsilon_m. \tag{10}$$

where ϵ_i is the usual own-price elasticity, ϵ_i^h is the Hicksian (or income-compensated or pure substitution) elasticity, and ϵ_m is income elasticity.

This equation is useful because you may have estimates of some of these elasticities and want to know others, which you can get from equation (10).

F. **Roy's Identity** shows how you can recover the ordinary (Marshallian) demand function from the indirect utility function $v(\mathbf{p}, m)$. It says

$$x_i^*(\mathbf{p}, m) = \frac{-\frac{\partial v}{\partial p_i}}{\frac{\partial v}{\partial m}} \tag{11}$$

See Varian p. 106-8 for three (!) proofs and some general remarks.

G. **Shepherd's Lemma** tells us how to recover the Hicksian (income compensated) demand functions from the expenditure function. (Recall an analogous expression by the same name on the supply side.) The demand side version is

$$h_i^*(\mathbf{p}, u) = \frac{\partial e(\mathbf{p}, u)}{\partial p_i}. \tag{12}$$

H. **Elasticities identity.** There is a useful formula that connects the values of the various demand elasticities. It is based on a formula discovered by the mathematician Leonhard Euler (1707-1783). The formula applies to homogeneous functions,

i.e., functions that for all $\mathbf{y}$ and all $a > 0$ satisfy the identity

$$f(a\mathbf{y}) = a^k f(\mathbf{y}) \tag{13}$$

for some nonnegative integer k. For example, Cobb-Douglas functions are homogeneous of degree k = sum of exponents. Euler showed that any such function can be written

$$y_1 \frac{\partial f}{\partial y_1} + ... + y_n \frac{\partial f}{\partial y_n} = k f(y_1, ..., y_n). \tag{14}$$

It is easy to verify (14) if you wish, simply by totally differentiating both sides of equation (13) with respect to a and evaluating at $a = 1$.

Note that Marshallian demand functions $x_i^*(p_1, ..., p_n, m)$ are homogeneous of degree 0: doubling all prices and income, for example, will have no effect on the optimal quantities of goods purchased, as you can see by inspecting the budget constraint. Applying (14) with $k = 0$ to $x_i^*(p_1, ..., p_n, m)$, we get

$$p_1 \frac{\partial x_i^*}{\partial p_1} + ... + p_n \frac{\partial x_i^*}{\partial p_n} + m \frac{\partial x_i^*}{\partial m} = 0. \tag{15}$$

Denote price elasticities by $e_{ij} = \frac{\partial x_i^*}{\partial p_j} \frac{p_j}{q_i}$ — if $i = j$ this is called the own-price, otherwise the cross-price elasticity of demand — and income elasticity by $\eta_i = \frac{\partial x_i^*}{\partial m} \frac{m}{q_i}$. Now divide all terms in (15) by the quantity demanded $q_i = x_i^*$ to obtain the desired identity

$$e_{i1} + ... + e_{in} + \eta_i = 0, \quad i = 1, ..., n. \tag{16}$$

That is, for any good i, the sum of its income elasticity and own price and all cross price elasticities are zero! You can impose this constraint directly on estimated log-linear demand functions to obtain more accurate results. Another way to

think of it is that own price elasticity is interesting to the extent that it differs from -1.0, cross price from from 0 and income elasticity from 1.0.

I. Moneysworth demand. See "Tractable consumer choice," by D. Friedman, Daniel and J. Skovics, *Theory and Decision* 79:2 pp.333–358 (2015).

- Recall that (aside from corner solutions) Marshallian demand is defined by the FOCs in (5). All except the last say that the marginal utility for each good is equal to λ times the price of that good. The last equation is the budget constraint.
- The equations for Hicksian demand are exactly the same except that the budget constraint is replaced by a utility constraint, $u(x_1, x_2, ...) = u_o$.
- Moneysworth demand also uses the same equations except that the budget constraint is replaced by an equation that says λ is the marginal value of keeping some cash for later purchases.
- The article above argues that this version of demand is simpler than either the Marshallian or Hicks versions. It gets rid of income effects, ties nicely to sequential decisions (we will study those later in the course), can handle indivisible goods (where you can't continuously adjust the amount) and liquidity constraints.
- The article also argues that moneysworth demand is more intuitive and realistic. People usually compare the satisfaction they get per unit of goods purchased to the value they expect to get for their money elsewhere. This can lead to behavior like money illusion and house-money effects that is realistic but inconsistent with Marshallian demand theory.

- As a practical matter, when you estimate an empirical demand function, the moneysworth perspective might help you find better proxies for the income variable.

J. Corners and notches. These notes emphasized interior solutions to the standard (or dual) consumer choice problem. Of course, a choke price p_i^c represents a point above which $\mathbf{x}^*$ is at the $x_i = 0$ corner (or face).

Sometimes additional constraints on consumer choice can be analyzed using only slight modifications. For example, the opportunity set is a natural modification of the budget set when there is a ceiling or floor on the amount purchased, or a subsidy or tax or matching grant.

Ex: Opportunity set with rationing (ceiling).

5. Notes on Risky Choice

Much of the material is covered in Varian Chapter 11. The Appendix below expands on some technical points.

1. Risk vs uncertainty.

   The opportunities considered so far ("bundles" of consumption goods) are riskless in the sense that you get to consume exactly what you choose. But some choices, especially in finance, are not like that. Your choices affect what you can consume, but you can't know exactly what you will get.

   In interesting situations you don't even know all the possible consequences of your choices, or the probabilities of some of the consequences; early 20th century economist Frank Knight referred to this as **uncertainty**. We will focus on the more tractable situations of what Knight called **risk**, where you know all possible consequences and their probabilities of each choice.

   Beyond the scope of this course lies research in the middle ground between Knightian uncertainty and risk. This includes Leonard Savage's work in the mid-20th century on subjective probabilities, recent behavioral work on ambiguity, and recent speculative forays on known versus unknown unknowns.

2. Risk as Lottery choice.

   For the rest of this course we will focus on cases where probabilities can be found for each possible outcome. The next unit shows how to use Bayes Theorem to compute probabilities from new information and previous beliefs. For now, we will assume that the probability $p_i \in [0, 1]$ is known for each possible outcome m_i.

   It is convenient to summarize each possible outcome as an increase (or possibly decrease) in purchasing power, relying on the standard theory of the previous unit that describes which bundles a person would choose given a total budget m and the corresponding level of satisfaction u. So here we will assume that each possible outcome m_i

is a real number, representing either total purchasing power (wealth) or, more often, an increase or decrease in purchasing power.

Thus we will regard choice as among a set of risky opportunities, each of which is described by possible monetary outcomes $m_i \in R$ with associated probabilities p_i. In some applications we take the range of outcomes to be an interval, in which case probabilities are given by a density function. For now, however, we assume that there are only a finite number of possible outcomes $i = 1, ..., k$ from any particular opportunity, so $\sum_{i=1}^{k} p_i = 1$. Economists refer to such well-defined risky opportunities as **lotteries**.

3. Modeling risky choice: Bernoulli rediscovered.

   How should a person choose among risky opportunities, i.e., among lotteries? The standard theory goes back to Daniel Bernoulli (1738); in the 1950s it was dusted off and made mainstream by researchers including John von Neumann, Leonard Savage, Ken Arrow, Milton Friedman and Harry Markowitz.

   The basic idea is that the decision maker has a utility function defined over all possible monetary outcomes, and that she rationally will choose the opportunity that maximizes utility on average. To explain, it is helpful to review some statistical and economic ideas.

   - Let L denote some lottery with outcomes $m_1, ..., m_k$ and associated probabilities $p_1, ..., p_k$. Then its **expected value** is $EL = \sum_{i=1}^{k} p_i m_i$, and its **variance** is $Var[L] = E(m - EL)^2 = \sum_{i=1}^{k} p_i (m_i - EL)^2$.
   - The expected value (sometimes denoted μ_L instead of EL) should be interpreted as the average monetary outcome, weighted by probability. Variance (sometimes denoted at σ_L^2 instead of $Var[L]$) is a measure of dispersion of monetary outcomes. Variance (or its square root, σ_L, called the **standard deviation**) is commonly used to quantify the riskiness of a lottery.
   - Suppose that the decision maker has the utility function $u(m)$, where u is a

differentiable increasing function, so $u' > 0$. Then the **expected utility** of lottery L is $E_L u = \sum_{i=1}^{k} p_i u(m_i)$.

- Suppose that the decision maker has to choose among a set of lotteries. According to **expected utility theory**, she will choose so as to maximize expected utility.

Here is a very simple example. The decision maker has utility function $u(m) = m^{0.6}$ defined for $m \geq 0$. She must choose either lottery R or lottery S. Lottery R pays \$10 if a fair coin comes up heads and \$0 otherwise. Thus $ER = 10 \cdot 0.5 + 0 \cdot 0.5 = \$5.00$ and $Var[R] = E(m - ER)^2 = (10-5)^2 \cdot 0.5 + (0-5)^2 \cdot 0.5 = 25$, so $\sigma_R = \sqrt{25} = \$5.00$. Lottery S pays \$4.00 for sure. Hence $ES = 4 \cdot 1.0 + 0 \cdot 0 = \$4.00$ and $Var[R] = E(m - ER)^2 = (4-4)^2 \cdot 1.0 + (0-4)^2 \cdot 0.0 = 0$, so $\sigma_S = \sqrt{0} = \$0.00$.

The decision maker's expected utility for lottery R is

$$E_R u = \sum_{i=1}^{k} p_i u(m_i) = p_1 u(m_1) + p_2 u(m_2) = .5 \cdot 10^{0.6} + 5 \cdot 0^{0.6} \approx .5 \cdot 3.98 = 1.99,$$

while $E_S u = 1.0 \cdot 4^{0.6} \approx 2.30$. Since $2.30 > 1.99$, expected utility theory predicts that she will pick S and not R. On the other hand (check this yourself) if another decision maker had utility function $w(m) = m^{0.9}$ and faced the same opportunities, then he would make the opposite choice.

4. Cardinal utility.

Notice that the last utility function w can be obtained from the previous utility function by means of a monotone increasing transformation: $w(m) = h(u(m)$ where $h(x) = x^{1.5}$. The theory used in previous section said that preferences (over bundles) are ordinal and thus not changed by applying such a transformation. What gives?

For several decades, economists criticized expected utility theory because it is not ordinal, but they eventually got over it. When it comes to evaluating lotteries, it is not enough to say that, for example, \$10 is better than \$4 which is better than \$0. Ordinal is not enough; you really have to know *how much* better. On the other hand, it doesn't affect comparisons of expected utility if you add or subtract a fixed amount

from every monetary outcome, or if you change the scale (say from dollars to cents, or to a foreign currency at fixed exchange rates).

The technical jargon is that risk preferences are **cardinal**, not ordinal, and that the utility functions we use here are unique only up to a positive affine transformation. It is straightforward to check that if lottery L1 has higher expected utility than L2 using utility function u, then it still has higher expected utility using utility function $v(m) = \alpha u(m) + b$ as long as $\alpha > 0$.

5. Useful utility functions.

   (a) The function family $f(m|r) = m^{1-r}/(1-r)$ is called **CRRA with parameter** r. We just saw two special cases, for $r = .4$ and $r = .9$. (Actually, we dropped the constant factor $1/(1-r)$, but that doesn't matter, as explained in the previous paragraph). This family is often used in macroeconomics, where the parameter r is chosen so as to align the model with data.

   Using L'Hospital's rule, you can show that in the limiting case $r \to 1$, the function f takes the form $f(m|1) = \ln(m)$, the function that Daniel Bernoulli originally proposed in 1738 !

   (b) The function family $u(m|a) = 1 - e^{-am}$ is called **CARA with parameter** $a > 0$. It is sometimes used in applied work, where the parameter a is fitted to the data. A higher value of a is interpreted as greater risk aversion, or more cautious preferences, as we will see next.

6. Measuring Risk Aversion.

   Given a twice continuously differentiable utility function u, the *coefficient of absolute risk aversion* at monetary outcome $m \in (-\infty, \infty)$ is

$$A(m) = \frac{-u''(m)}{u'(m)}. \tag{1}$$

   It is straightforward to verify that $A(m) = a$ in the CARA function family. That is why it has its name: CARA is an acronym for constant absolute risk aversion.

The *coefficient of relative risk aversion* at $m > 0$ is

$$R(m) = \frac{-u''(m)}{u'(m)} m = mA(m). \tag{2}$$

It will not surprise you to hear (but check this anyway) that $R(m)$ is constant for all functions in the CRRA family, including $\ln m$. Can you now decipher the CRRA acronym?

You should also check that the functions $A(m)$ and $R(m)$ are unaffected by positive affine transformations, that is, they are the same for $v(m) = \alpha u(m) + b$ as they are for u if $\alpha > 0$. So these are valid cardinal measures of risk preferences.

7. Intuition about A and R.

   Suppose that $w(m) = \alpha m + b$ is linear and upward sloping, $\alpha > 0$. Then w represents the same cardinal preferences as $u(m) = m$. Of course, expected utility for u is exactly the same thing as expected value, so the same is true for w. Thus according to expected utility theory, a person with increasing linear utility will always choose the lottery with highest expected value, irrespective of variance. Such a person is said to be **risk-neutral**.

   A risk-neutral person has $u''(m) = 0$, and so by equations (1) and (2), such a person has $A(m) = R(m) = 0$.

   Higher values of A and R indicate greater aversion to risk. Why? Look at equations (1) and (2) again. They are sign-adjusted measures of concavity, appropriately normalized, and greater concavity means less willingness to accept risk. To spell this out,

   - recall that $u'' < 0$ for a concave function, so the numerators are $-u'' > 0$.
   - The denominators u' normalize u'' so that positive affine transformations have no effect, as you checked with u and v in the previous item.
   - Greater normalized curvature implies a larger drop in utility for a risky lottery relative to a non-risky lottery with the same expected value.

The last point is illustrated in Figure 1. There would be no gap between $u(EL)$ and $E_L u$ if the utility function were linear, and a very large gap if the utility function were tightly curved in the neighborhood of EL. The next item spells out the connection between risk and utility.

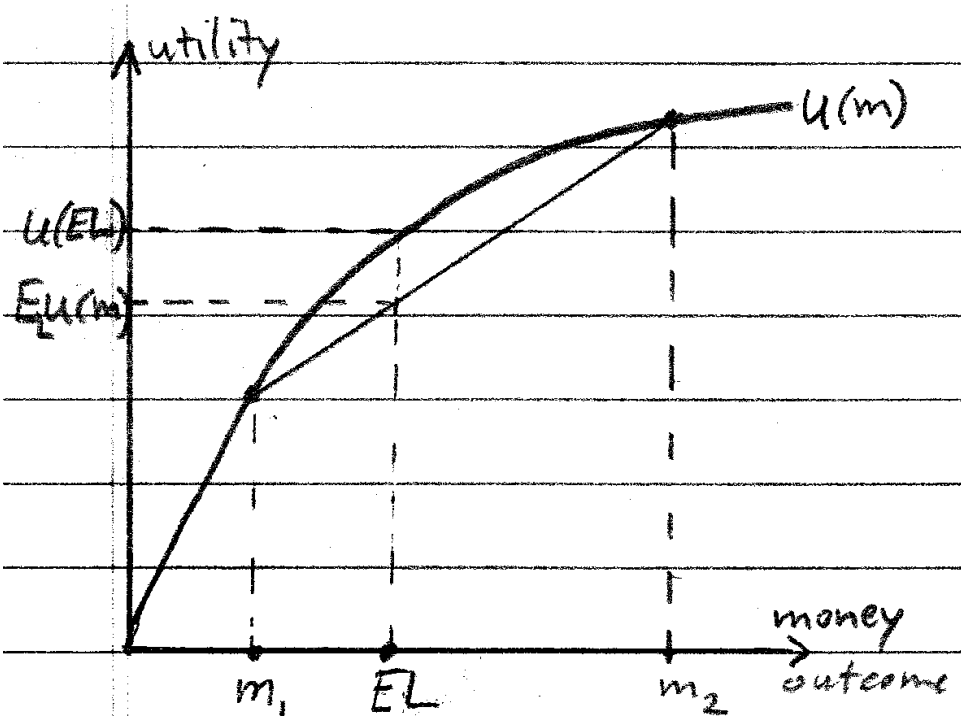


Figure 1: Expected Utility. Here the lottery prizes are $m_1 = 20$ and $m_2 = 90$ with probabilities $p_1 = .3$ and $p_2 = .7$; the expected value is $EL = 69$. The expected utility $E_L u = p_1 u(m_1) + p_2 u(m_2)$ is the height of the point above EL on the line segment connecting the points $(m_1, u(m_1))$ and $(m_2, u(m_2))$ on the utility curve. The gap between $u(EL)$ on the u-curve and the expected utility $E_L u$ on the line segment is larger the greater the degree of concavity (curvature) in the utility function u.

8. Mean and Variance.

By Taylor's Theorem, any smooth (continuously differentiable 3 times) utility function u can be expanded at any point z in its domain as a quadratic function plus remainder:

$$u(z+h) = u(z) + u'(z)h + \frac{1}{2}u''(z)h^2 + R^3(z,h), \tag{3}$$

where $R^3(z,h) = \frac{1}{6}u'''(y)h^3$ for some point y between z and $z+h$.

In equation (3), set $z = EL$ and $m = z + h$, and take the expected value of both sides. The second (linear) term disappears because $Eh = E(m - EL) = EL - EL = 0$. Since $Eh^2 = Var[L]$ the equation becomes

$$E_L u(m) = u(EL) + \frac{1}{2}u''(EL)Var[L] + R^3. \tag{4}$$

That is, expected utility of the lottery is equal to the utility of the mean outcome, plus a term proportional to the variance of the lottery and to the second derivative of u evaluated at the mean of the lottery, plus a remainder term.

- The second term in equation (4) is key. It says that variance reduces expected utility to the extent that u is concave, as measured by (unnormalized) $A(m)$. Otherwise put, a person with higher A will be more averse to variance than another rational person with A closer to zero. If $A(m) = 0$ for all m, then the Bernoulli function is linear and that person is risk neutral.
- We can ignore the remainder term if either

  (a) h is small because all likely outcomes m are near EL, or

  (b) $u'''(EL)$ is small because u'' is almost constant in the neighborhood of EL.

  In other words, the approximation using the first two terms is reliable if either the risks are small, or the utility function is nearly quadratic. A separate argument shows that the approximation is exact if the probability distribution is Normal (aka Gaussian).
- Using a suitable affine transformation and ignoring the remainder term, expected utility takes the form $EL - cVar[L]$, or in other notation $\mu_L - c\sigma_L^2$, where c is proportional to the coefficient of absolute risk aversion.
- A lot of the finance literature assumes directly that utility takes this mean-variance form. In many important cases it is a good approximation, but it can be misleading for highly risky non-Normal lotteries. See homework problem 3.2.

9. Justifications for EUT.

   There are two reasons to think that EUT should work, i.e., that a person would have some particular utility function $u(m)$ and, when faced with a set of risky opportunities, would choose one with highest expected utility.

   (a) EUT says that people choose what, on average, will make them happiest. This seems reasonable, and was Bernoulli's original justification, back in 1738.

   (b) If you write down a general preference relation over lotteries (instead of ordinary bundles), the same basic properties we used in the last unit ensure that it can be represented by a utility function U defined on the space of lotteries. This will not necessarily take the form $U(L) = E_L u$ for some utility function u defined on real numbers (monetary outcomes). But a beautiful theorem (developed in the 1950s by the economists mentioned earlier) shows that it does take that form if a few additional, very reasonable, properties are assumed.

   Thus a second reason for EUT is that it is the logical consequence of some very sensible axioms (like continuity, transitivity, and other consistency requirements) about how people make risky choices.

   This Expected Utility Theorem is explained in the Appendix below. (The function u is called a Bernoulli function there to avoid confusing it with the utility function U.) See also Varian pp. 173-176, and many other references.

10. Some caveats

    - Unless you are trained to do so, it seems unnatural to actually calculate expected utilities when making a risky choice decision.
    - EUT should be understood to say that, if a person is rational and consistent in the face of risk, her choices can be nicely summarized by finding a utility function such that her actual choices maximize the expectation of that function over available opportunities.

- Unfortunately, lots of empirical research, in the lab and in the field, has not uncovered utility functions that enable EUT to make good predictions of actual risky choices by most people.
- Behavioral economists have used Prospect Theory (Kahneman and Tversky, 1979) to account for many discrepancies. PT generalizes EUT by (a) categorizing m's as either gains or losses relative to some reference point, (b) allowing the utility function u to be concave in gains, convex in losses, and kinked at the reference point, and (c) allowing for distorted probabilities, especially overweighting the probability of rare large gains or losses.
- A recent book, *Risky Curves* by D. Friedman et al. (Routledge, 2013), argues that PT has so many free parameters that, although it can account well for many data sets after the fact, it shows little or no improvement relative to EUT in predicting what people will do in new settings. The book argues that, as far as prediction is concerned, sophisticated versions of risk-neutral expected utility (taking into account indirect effects) may offer the best predictions currently available.

Bottom line. In your instructor's opinion, EUT is good normative theory (i.e., in saying what you *should* do), and is an essential building block of modern finance. But you should be cautious in using it to predict what ordinary people will do.

Appendix: Some formalities.

Basic definitions. A *lottery* $L = (M, P)$ is a finite list of monetary outcomes $M = \{m_1, m_2, ..., m_k\} \subset \Re$ together with a corresponding list of probabilities $P = \{p_1, p_2, ..., p_k\}$, where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. The symbol $\Re$ denotes the real numbers $(-\infty, \infty)$. The *space of all lotteries* is denoted $\mathcal{L}$.

The *expected value* of lottery $L = (M, P)$ is $EL = \sum_{i=1}^{k} p_i m_i$.

A utility function over monetary outcomes (henceforth called a *Bernoulli function*) is a strictly increasing function $u : \Re \to \Re$. A *utility function over lotteries* is a function $U : \mathcal{L} \to \Re$.

Given a Bernoulli function u, the *expected utility* of lottery $L = (M, P)$ is $E_L u = \sum_{i=1}^{k} p_i u(m_i)$.

Preferences $\succeq$ over any set refer to a complete and transitive binary relation. A utility function U *represents* preferences $\succeq$ if $x \succeq y \iff U(x) \geq U(y)$ for all x, y in that set, where the symbol " $\iff$ " means "if and only if." In particular, a utility function $U : \mathcal{L} \to \Re$ represents preferences $\succeq$ over $\mathcal{L}$ if

$$L \succeq L' \iff U(L) \geq U(L')$$

for all $L, L' \in \mathcal{L}$.

Expected utility theorem. Preferences $\succeq$ over $\mathcal{L}$ have the *expected utility property* if they can be represented by a utility function U that is the expected value of some Bernoulli function u. That is, there is some Bernoulli function u, such that for all $L, L' \in \mathcal{L}$, we have $L \equiv (M, P) \succeq (M', P') \equiv L' \iff$

$$U(L) \equiv E_L u \equiv \sum_{i=1}^{k} p_i u(m_i) \geq \sum_{i=1}^{k} p'_i u(m'_i) \equiv E_{L'} u \equiv U(L'). \tag{5}$$

It might seem that preferences with the expected utility property are quite special, and indeed they are. For example, their indifference surfaces are parallel and flat. Thus preferences over for lotteries with only $k = 3$ monetary outcomes have indifference curves on $\mathcal{L}$

(represented as the probability simplex, here a triangle) that are all straight lines with the same slope.

The expected utility theorem (EUT) is therefore surprising. It states that preferences over lotteries that satisfy a seemingly mild set of conditions will automatically satisfy the expected utility property, and thus be representable via a Bernoulli function.

Over the decades since the original results of Von Neumann and Morgenstern, many different sets of conditions have been shown to be sufficient. Here we mention the set used in a leading textbook, Mas-Colell et al. [2010]. It consists of four axioms that an individual's preferences $\succeq$ on $\mathcal{L}$ should satisfy:

1. Rationality: Preferences $\succeq$ are complete and transitive on $\mathcal{L}$.

2. Continuity: The precise mathematical expressions are rather indirect (they state that certain subsets of real numbers are closed sets), but they capture the intuitive idea that U doesn't take jumps on the space of lotteries. This axiom rules out lexicographic preferences.

3. Reduction of Compound Lotteries: Compound lotteries have outcomes that are themselves lotteries in $\mathcal{L}$. By taking the expected value, one obtains the *reduced* lottery, a simple lottery in $\mathcal{L}$. The axiom states that the person is indifferent between any compound lottery and the corresponding reduced lottery.

4. Independence: Let $L, L', L'' \in \mathcal{L}$ and $\alpha \in (0, 1)$. Suppose that $L \succeq L'$. Then $\alpha L + (1-\alpha)L'' \succeq \alpha L' + (1-\alpha)L''$. "In other words, if we mix two lotteries with a third one, then the preference ordering of the resulting two mixtures does not depend upon (is independent of) the particular third lottery used." (Mas-Colell et al. p. 171).

Theorem 1 (EUT)**.** *Let preferences $\succeq$ on $\mathcal{L}$ satisfy axioms 1-4 above. Then $\succeq$ has the expected utility property, i.e., there is a Bernoulli function u such that (5) holds.*

As Mas-Colell et al. point out, all four axioms seem innocuous. Someone who cares only

about the ultimate monetary payoffs and whose calculations are not affected by indirect ways of stating the probabilities will satisfy the third axiom. For example, such a person would be indifferent between the compound lottery "get 0 with probability 0.5 and with probability 0.5 play the lottery that pays 10 with independent probability 0.5 and 0 otherwise," and the reduced lottery "get 0 with probability 0.75 and 0 with probability 0.25." The fourth axiom enforces a degree of consistency by requiring that preference rankings over lotteries are not changed by nesting each of those lotteries within a generic compound lottery. The first two axioms are even less controversial or problematic.[1]

For a proof of the EUT, see Mas-Colell et al. and the references cited therein. Here is a sketch of how the function u can be constructed for given preferences. Denote by m_+ and m_- the maximum and minimum monetary outcomes in the lottery. Set $u(m_+) = 1$ and $u(m_-) = 0$. Consider any other monetary outcome m, and the set of lotteries $\{([m_+, m_-], [p, (1-p)]) : p \in [0, 1]\}$. For $p = 1$ the lottery is preferred to m and for $p = 0$ the outcome m is preferred to the lottery. Using the continuity axiom, one can show that for some intermediate p^* the person is indifferent between m and that lottery. Set $u(m) = p^*$. Then use the other axioms to verify that the Bernoulli function so constructed indeed represents the given preferences.

Notice that this construction of the Bernoulli function suggests an empirical procedure: vary the probabilities on best and worst outcomes to try to find a person's point of indifference. There are many variations on this theme; one of the currently more popular is the Multiple Price List scheme introduced by Holt and Laury (*American economic review*, 2002).

The rest of the material in these notes will not be covered in Econ 200, but students taking a course in Finance may (or may not!) find it helpful.

[1]On the other hand, the EUT's conclusion is quite strong, and not consistent with some actual choice data. One response is to accommodate some of the anomalous data by weakening the axioms, usually the third or fourth. For an extensive and skeptical review of these matters, see my recent book *Risky Curves* (Routledge, 2014) coauthored with M. Isaac, D. James and S. Sunder.

0.1 Basic Definitions

The *return* k on an asset over a given period is defined in terms of the cash flow x received during that period and the beginning and end prices P_0 and P_1 as follows:

$$k = \frac{P_1 - P_0 + x}{P_0}. \tag{6}$$

Suppose that the return in situation (or scenario) s is k_s, and that each possible situation $s = 1, ..., S$ has probability $p_s > 0$, where $\sum_s p_s = 1$. Then the *expected return* is

$$Ek = \sum_{s=1}^{S} p_s k_s. \tag{7}$$

An important special case is with equally-weighted historical data. Each observation receives equal probability weight $p_s = 1/S$.

The *variance* is

$$\text{VAR } k = \sigma_k^2 = E(k - Ek)^2 = \sum_{s=1}^{S} p_s (k_s - Ek)^2. \tag{8}$$

The square root of the variance, $\sigma_k = \sqrt{\text{VAR } k}$ is called the *standard deviation* of the return, or the *volatility* of the asset. It is the most popular measure of risk.

The *covariance* between returns k and h on two different assets is

$$\text{COV}(k, h) = \sigma_{kh} = E(k - Ek)(h - Eh) = \sum_{s=1}^{S} p_s (k_s - Ek)(h_s - Eh). \tag{9}$$

The *correlation* ρ_{kh} is covariance normalized by the volatilities. Thus $\rho_{kh} = \text{COV}(k, h)/(\sigma_k \sigma_h)$. Standard theorems in mathematics demonstrate that $-1 \leq \rho_{kh} \leq 1$.

Consider a portfolio $x = (x_k, x_h)$ with initial dollar value x_k in asset k and initial value x_h in asset h. The numbers x_k, x_h are called *positions* or *exposures*. Usually we have $x_k > 0$, which is called a *long* position, but a *short* position $x_k < 0$ sometimes occurs, e.g., if you borrow an asset. The expected return on portfolio x with total initial investment $x_T = x_k + x_h$ is $(x_k/x_T)Ek + (x_h/x_T)Eh$. It is the obvious weighted average of the returns on the two assets.

0.2 Diversification

Portfolio variance is more complicated and interesting. First note that the definitions imply that:

- variance $\text{VAR } ak = a^2 \text{VAR } k$ increases in the square of the position a
- volatility $\sigma_{ak} = |a|\sigma_k$ increases linearly in the absolute position
- the portfolio with unit positions in two assets has variance

$$\begin{aligned} \text{VAR } (k+h) &= E[(k+h) - E(k+h)]^2 = E(k - Ek)^2 + 2E(k - Ek)(h - Eh) + E(h - Eh)^2 \\ &= \text{VAR } k + 2\text{COV}(k,h) + \text{VAR } h. \end{aligned} \tag{10}$$

From the first and third bullet items above, it follows that the variance on the portfolio x above is $x_k^2 \text{VAR } k + 2x_k x_h \text{COV}(k,h) + x_h^2 \text{VAR } h$.

This piece of mathematics has a very important implication. It shows that risk can be reduced substantially by diversification. To see this clearly, suppose that the two assets have the same variance, say $\text{VAR } k = \text{VAR } h = 100$, and consider various portfolios where the positions add to 10, say $x_k = a$ and $x_h = 10 - a$. Clearly if all 10 is put in either asset and the other is left out, then the portfolio volatility is $|10|\sigma_k = 10 * 10 = 100$. As discussed in class (using diagrams in (σ_k, Ek) space), there are three special cases of interest:

1. if $\rho_{kh} = 0$, then the portfolio volatility is $\sqrt{a^2\sigma_k^2 + (10-a)^2\sigma_h^2} < 100$. For an equal-weight portfolio $a = 5$, for example, the portfolio volatility is only $\sqrt{25 * 100 + 25 * 100} = 50\sqrt{2} \approx 70.7$, almost a 30% reduction in volatility or riskiness.

2. if $\rho_{kh} = -1$, then risk can be eliminated entirely. Now $\text{COV}(k,h) = \rho_{kh}\sigma_k\sigma_h = -1 * 10 * 10 = -100$. For $a = 5$, the portfolio volatility is $\sqrt{5^2\sigma_k^2 + 2 * 5 * 5\text{COV}(k,h) + 5^2\sigma_k^2} = \sqrt{2500 - 5000 + 2500} = 0$. The idea is that when

$\rho = -1$, any fluctuation in the first asset is exactly offset by the fluctuation in the other asset, and risk is eliminated.

3. if $\rho_{kh} = 1$, then there is no diversification effect and all portfolios with the same total position have the same risk. The mathematical reason is that now $\text{Cov}(k, h) = \sigma_k \sigma_h$. In the example, the portfolio volatility is $\sqrt{a^2\sigma_k^2 + 2 * a * (10 - a)\sigma_k\sigma_h + (10 - a)^2\sigma_k^2}$ $= 10\sqrt{a^2 + 2 * a * (10 - a) + (10 - a)^2} = 10 * \sqrt{100} = 100$ for any a.

The take-home point here is that there is a reduction in risk when a given initial investment is spread among several assets whose returns are not perfectly correlated. For positive positions, the diversification effect is stronger when the correlation is smaller (or more negative). The effect strengthens as the fraction of the investment in one asset increases from zero, and eventually weakens as the fraction approaches one.

Exercises.

1. Show that if the volatilities of two assets are not equal, you can still find a riskless portfolio when $\rho = -1$, with positions related to the relative volatilities of the two assets.

2. Also show that risk can be eliminated when $\rho = 1$ by taking a short position in one asset and a long position in the other.

0.3 Marginal Risk in a Portfolio

The rest of these notes concern portfolios involving N risky assets with returns k_i for $i = 1, ..., N$ and covariance matrix $C = ((\sigma_{ij}))$. (Note that the diagonal elements $\sigma_{ii} = \sigma_i^2$ are the variances of the risky asset, and the off-diagonal elements are covariances as defined earlier.) The positions are denoted x_i and the total investment is $x_T = \sum_{i=1}^{N} x_i$. The portfolio shares are $a_i = (x_i/x_T)$. The reasoning in the previous section shows that the expected return on

portfolio x is

$$E[x] = a \cdot Ek = \sum_{i=1}^{N} a_i Ek_i, \tag{11}$$

the weighted average expected return on the N assets, with weights given by the portfolio shares. The earlier reasoning also shows that the variance of the portfolio value is

$$\textsc{Var}[x] = x \cdot Cx = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \sigma_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \sigma_i \sigma_j \rho_{ij}, \tag{12}$$

where $\rho_{ij} = \textsc{Cov}(k_i, k_j)/(\sigma_i \sigma_j)$ is the correlation between the returns on assets i and j. The portfolio volatility is denoted $\sigma_x = \sqrt{\textsc{Var}[x]}$.

The question is: Holding constant the overall investment, how does the portfolio volatility change as the the position in one particular asset i increases? Or more briefly, what is the marginal risk of asset i in portfolio x?

The question is key to portfolio analysis, and its answer requires a little linear algebra and vector calculus. Let $x = (x_1, ..., x_N)$ be the original portfolio and let $y = \alpha x_T e^i + (1-\alpha)x$ be the shifted portfolio, where $\alpha \geq 0$ is the proportional amount of the shift in direction $e^i = (0, ..., 0, 1, 0, ..., 0)$, the unit portfolio with only asset i. The question is how fast portfolio volatility σ_y changes as α increases from zero. The question and its answer are captured in:

Lemma 1. For portfolio y as just defined, we have

$$\left.\frac{d\sigma_y}{d\alpha}\right|_{\alpha=0} = \frac{-\sigma_x^2 + \sigma_{ix}}{\sigma_x}. \tag{13}$$

That is, portfolio volatility increases at a rate proportional to the covariance $\sigma_{ix} = x_T e^i \cdot Cx$ of asset i with the rest of the portfolio. The other term $-\sigma_x^2/\sigma_x = -\sigma_x$ in (13) simply reflects the reduced size of the rest of the portfolio when the asset i share increases.

The following proof can be skipped by readers who do not enjoy such things.

Proof. From equation (12) and the definition of y we have

$$\textsc{Var}[y] = y \cdot Cy = (\alpha x_T e^i + (1-\alpha)x) \cdot C(\alpha x_T e^i + (1-\alpha)x) \tag{14}$$

$$= (1-\alpha)^2 x \cdot Cx + \alpha^2 x_T^2 e^i \cdot Ce^i + 2\alpha x_T (1-\alpha) e^i \cdot Cx, \tag{15}$$

where the last expression uses the fact that $e^i \cdot Cx = x \cdot Ce^i$ since the covariance matrix is symmetric. Hence

$$\left.\frac{d\text{VAR}[y]}{d\alpha}\right|_{\alpha=0} = -2(1-\alpha)x \cdot Cx + 2\alpha x_T^2 e^i \cdot Cx + 2(1-2\alpha)x_T e^i \cdot Cx \,|_{\alpha=0} \tag{16}$$

$$= -2x \cdot Cx + 2x_T e^i \cdot Cx = -2\sigma_x^2 + 2\sigma_{ix}. \tag{17}$$

Marginal risk now can be calculated using the chain rule of ordinary calculus:

$$\left.\frac{d\sigma_y}{d\alpha}\right|_{\alpha=0} = \left.\frac{d\sqrt{\text{VAR}[y]}}{d\alpha}\right|_{\alpha=0} = \frac{1}{2}[\text{VAR}[y]]^{-1/2} \left.\frac{d\text{VAR}[y]}{d\alpha}\right|_{\alpha=0} = \frac{-2\sigma_x^2 + 2\sigma_{ix}}{2\sigma_x}, \tag{18}$$

which immediately simplifies to the desired expression. $\diamond$

0.4 The Capital Asset Pricing Model (CAPM)

CAPM is no longer unchallenged orthodoxy but it still is the central model in modern financial theory. Its main result is that the expected return of each risky asset i is the risk-free return k_F, plus a risk premium which is the market price of risk RP_M times the quantity of risk β_i for that asset. That is, we have the

Security Market Line Theorem. If conditions B1, B2, M1 and M2 below hold, then the return on each asset i satisfies

$$Ek_i = k_F + \beta_i \text{RP}_M \tag{19}$$

where $\text{RP}_M = E[M] - k_F$ is the expected return on the Market portfolio—the portfolio consisting of all assets traded in financial markets—in excess of the risk-free return, and $\beta_i = \sigma_{iM}/\sigma_M^2 = \rho_{iM}\sigma_i/\sigma_M$ is the normalized covariance of asset i with the market portfolio.

The conclusion follows from various sets of assumptions. One short, convenient and transparent set of behavioral (B) and market (M) assumptions is as follows.

- (B1) Investors are rational and care only about portfolio expected return (+) and volatility (-).

- (B2) They have homogeneous beliefs about asset returns Ek and covarariances C.
- (M1) A financial market lasts just one period (from $t = 0$ to $t = 1$); it is competitive and frictionless (i.e., no market power and zero transactions costs) and is in equilibrium.
- (M2) All N risky assets are traded on the financial market plus one risk-free asset with return $k_F > 0$. The portfolio M consisting of all assets has expected return $E[M] > k_F$.

Here is a sketch of the proof of the SML Theorem; see a standard text for diagrams, etc. First plot in (σ, Ek)-space the volatilities and expected return for each risky asset in isolation. Using equations (3, 11, 12), note that the portfolios consisting of varying shares of two or more risky assets trace out parabolic arcs that fill a convex region of that space. Efficient risky portfolios (those with the largest expected return for given volatility, or smallest volatility for given expected return) lie along the Northwestern frontier of this region.

The Capital Market line (CML) has one endpoint at the risk-free point F with coordinates $\sigma = 0$ and $Ek = k_F > 0$, and passes through a point $T = (\sigma_T, E[T])$ of tangency to the Northwestern frontier. That is, the line satisfies the equation

$$E[x] = k_F + [(E[T] - k_F)/\sigma_T]\sigma_x \tag{20}$$

Since the market is frictionless, each investor can borrow or lend as much as he or she wants at the risk-free rate. Taking this into account, each efficient portfolio must lie on the CML, and by (B1) each investor will choose an efficient portfolio. Thus we have the

Markowitz Separation Theorem. Each investor's I's portfolio is of the form

$$w_I a_I F + w_I(1 - a_I)T, \tag{21}$$

that is, she lends some fraction a_I of her wealth w_I at the risk-free rate k_F (or borrows if $a_I < 0$) and invests the rest in the tangency portfolio.

The striking implication is that investors' risk preferences do not affect the mix of risky assets—all investors choose the same combination T. It's just that more risk averse investors

put more of their wealth into the riskless asset, and the least risk averse investors borrow at the riskless rate to buy more of the T portfolio, i.e., they use leverage. This striking implication, of course, is not exactly true in the real world, but some finance economists argue that the vast majority of investors (at least in terms of wealth) choose portfolios that are near T in (σ, Ek)-space.

The next step in the argument (due to Sharpe and Lintner in the 1960s) is that financial market equilibrium implies that the tangency portfolio T must be the same as the market portfolio M. To spell it out, the market demand for risky assets is the N-vector $w\tau$, where $w = \sum_I w_I(1-a_I)$ is the total wealth invested in risky assets, and $\tau = (\tau_1, ..., \tau_N)$ is the unit vector of assets that produces the point T in (σ, Ek)-space. The supply of risky assets is the N-vector $(s_1, ..., s_N)$. But in equilibrium, supply = demand, i.e., $s_i = w\tau_i$ for each asset i. Thus portfolios M and T have identical shares of the risky assets and therefore represent the same point in (σ, Ek)-space.

The SML (19) now follows from direct calculations. Let $y = \alpha e^i + (1-\alpha)M$ be a small shift in the market portfolio towards asset i along the efficient frontier. The slope of the frontier at M in (σ, Ek)-space, according to the implicit function theorem, is given by a ratio of derivatives, the numerator being $\frac{dE[y]}{d\alpha}\Big|_{\alpha=0}$ and the denominator being $\frac{d\sigma_y}{d\alpha}\Big|_{\alpha=0}$. By (11) the numerator is $\frac{d}{d\alpha}[\alpha Ek_i + (1-\alpha)E[M]]\big|_{\alpha=0} = Ek_i - E[M]$. The denominator is given by Lemma 1 of the previous section as $\frac{-\sigma_M^2+\sigma_{iM}}{\sigma_M}$. Since the CML is tangent to the frontier at M, the slope of the frontier just computed must be equal to the slope of the CML given by the bracketed term in (20), i.e.,

$$\frac{Ek_i - E[M]}{\frac{-\sigma_M^2+\sigma_{iM}}{\sigma_M}} = \frac{(E[M]-k_F)}{\sigma_M}. \tag{22}$$

Cross-multiplying we get

$$Ek_i - E[M] = (E[M]-k_F)\frac{-\sigma_M^2+\sigma_{iM}}{\sigma_M^2} = (E[M]-k_F)\frac{\sigma_{iM}}{\sigma_M^2} - E[M] + k_F. \tag{23}$$

Cancelling the $-E[M]$ terms on both sides of the equation we have exactly the expressions given in the Security Market Line Theorem.

6. Decision Theory

Not actually covered in Varian or most undergrad micro text; lots of good advanced books, but none at the right level. In order of decreasing accessibility, consider reading:
An Introduction to Bayesian Inference and Decision, by Robert L. Winkler (Holt, Reinhart and Winston, 1972).
Optimization in Economic Theory, Second Edition, by A.K. Dixit (Oxford, 1990).
The Analytics of Uncertainty and Information, by Hirshleifer and Riley (Cambridge University Press, 1992).
Optimal Statistical Decisions, by Morris DeGroot (McGraw-Hill, 1970).

I. Overview

A. Static optimization used so far neglects crucial complications such as:

1. You might have better information later, but delay is costly.
2. Your decision now may affect your opportunities (and information) later.

B. Decision theory and practice go back centuries and are still maturing, but saw major advances in the late 20th century.

C. Example applications:

1. Valuing real options, e.g., when to invest, or shut down or expand operations.
2. Valuing information and learning, e.g., when are consultants worth their fees?
3. Valuing financial options, and related questions central to modern finance.

II. Bayes Theorem

A. How should you revise your beliefs when new information arrives?

B. First breakthrough in Decision Theory, goes back to late 1700s (Rev. Bayes, Laplace).

C. Here's a basic guide, beginning with notation.

1. $m = a, b, c, ... \in M$ are the possible **messages**, or new pieces of information, that may arrive.
2. $s = A, B, C... \in S$ are the possible true **states** of Nature.
3. $p(m, s) \geq 0$ are the joint probabilities that m is observed and s is the true state. Note that $\sum_{s \in S} \sum_{m \in M} p(m, s) = 1$.
4. $p(m) = \sum_{s \in S} p(m, s) > 0$ is the probability of message m.
5. $p(s) = \sum_{m \in M} p(m, s) > 0$ is the probability of state s. Bayesians call it the **prior probability** of state s.
6. $p(m|s) = p(m, s)/p(s)$ is the conditional probability that m will be observed when the true state is s. Bayesians call it the **likelihood** of m given s.
7. $p(s|m) = p(m, s)/p(m)$ is the conditional probability that the true state is s when m is observed. Bayesians call it the **posterior probability** of s given m.

D. The goal is to properly compute the posterior probability after you observe a particular message m, when your prior beliefs were summarized by the $p(s)$'s. Bayes theorem tells you how to do that when you know the message likelihoods.

1. For example, suppose that you know that in the US, one person in a million with flu-like symptoms has zika, and there is a quick test that is 99.0% accurate in detecting zika when present (i.e., probability is 1% for both type I and type II errors).
2. Suppose your friend comes in with flu-like symptoms, and then gets a positive test result. Given that bad news, what is the probability that she actually has zika?

E. Here are three equivalent versions Bayes Theorem.

$$p(s|m) = \frac{p(m|s)}{\sum_{t \in S} p(m|t)p(t)} p(s) \quad (1)$$

$$\frac{p(s|m)}{p(t|m)} = \left[\frac{p(m|s)}{p(m|t)}\right] \left[\frac{p(s)}{p(t)}\right] \quad (2)$$

$$\ln[\text{posterior odds}] = \ln[\text{likelihood ratio}] + \ln[\text{prior odds}] \quad (3)$$

F. In words,

- the first equation says that the posterior state probability is proportional to the prior probability (with the proportion being the likelihood of the message given that state s, relative to all alternative states t).
- The second equation says that the posterior odds for any two states are equal to the prior odds times the likelihood ratio.
- The third equation just takes logs of the second equation.

G. The proof is simple, just unwinding definitions.

1. Begin with definition 7 above for $p(s|m)$, and insert def. 4 in the denominator.
2. Then rewrite the numerator $p(m, s) = p(m|s)p(s)$, using def 6.
3. Finally, again use def 6 for all summed terms in the denominator, $p(m, t) = p(m|t)p(t)$.
4. The final expression is exactly the first equation.
5. To get the second equation, simply divide the version of (1) for s by the version of (1) for another particular state t. The denominators in (1) cancel.
6. As already noted, (3) follows from (2) by taking logs.

H. Now let's solve the zika problem, and reassure your friend.

1. Let $s = A$ be the state of Nature where she actually has zika, and $s = B$ the state where she does not. So $S = \{A, B\}$.
2. Let $m = a$ be the positive test result, and $m = b$ the negative test result. So $M = \{a, b\}$.
3. Prior probabilities $p(A) = 0.000001$ and $p(B) = 0.999999$ are given, as are likelihoods $p(a|A) = 0.99$ and $p(b|B) = 0.99$. The error probabilities are $p(b|A) = p(a|B) = 0.01$.
4. Equation (1) says
   $p(A|a) = \frac{p(a|A)}{p(a|A)p(A)+p(a|B)p(B)}p(A) = \frac{.99}{(.99)(.000001)+(.01)(.999999)}(.000001) = 0.0000989.$

5. Equation (2) says that the posterior odds are
$\frac{p(A|a)}{p(B|a)} = \left[\frac{0.99}{0.01}\right]\left[\frac{0.000001}{.999999}\right] \approx [100][0.000001] = 0.0001.$
6. Using base 10 logs (any base will do!), equation 3 says the posterior log odds are about 2 + (-6) = -4.
7. Bayes Theorem tells us that the very strong prior (one in a million) dwarfs the pretty good message (about 100 to 1), so after the test your friend still has only about 1 chance in 10,000 of having zika.
8. Here's some intuition, in terms of frequencies. Out of a million folks with flu-like symptoms, on average only one actually has the disease but about 1% or 10,000 get a positive test result. Hence the posterior odds are about 1 in 10,000.

I. A moderately complicated finite problem is probably best done on a spreadsheet. Here's a good layout.

- The first column lists the true states $A, B, C, ... \in S$.
- The second column lists the given prior probabilities.
- The next $|M|$ columns list the given likelihoods $p(m|s)$ of each message $a, b, c... \in M$, one column for each message m.
- The rest of the spreadsheet calculates all other probabilities, using the formulas already introduced.
- The next set of $|M|$ columns lists the joint probabilities of each message-state pair, $p(m, s) = p(m|s)p(s)$, computed from preceding columns.
- Each column sum is the message probability $p(m) = \sum_{s\in S} p(m, s)$.
- The final set of $|M|$ columns lists the posterior probabilities $p(s|m) = p(m, s)/p(m)$, computed using a preceding column (and its sum).
- An example in excel can be found on the class website.

J. More on Bayes Theorem

1. If the message is independent of the state, then

- $p(m, s) = p(m)p(s)$ and
- Bayes Theorem tells us that the posterior is the same as the prior, i.e., $p(s|m) = p(s)$ for all $m \in M$.
- Such messages are uninformative, i.e., useless.

2. If you have continuous message spaces and/or continuous state spaces, then
   - the formulas apply to the density functions $f(m, s)$ etc replacing the discrete probabilities $p(m, s)$ etc.
   - You just have to work with states and messages where the relevant densities are continuous and positive.
   - Mathematically inclined readers who know epsilon-delta arguments should have no problem proving this by taking the limit of discrete approximations to f.
3. What if you get several different messages, one after another? Or run several different diagnostic tests?
   - Then just reapply Bayes theorem, where the new prior is the old posterior.
   - If messages are conditionally independent from each other (hopefully not independent of the true state!) then you can get a factor in (2) for each message, and
   - (3) says that you sum all the message log likelihoods (plus the log prior odds) to get the log posterior odds.
4. Does (3) remind you of maximum likelihood estimation (MLE)?
   - There is a connection. Think of each observation as a separate message, and of all possible parameter values as S, the states of Nature.
   - In standard MLE, the likelihoods are functions of the unknown parameters, and one can search parameter space to find parameter values that maximize the sum of log likelihoods.
   - In standard MLE, in effect the prior odds are set to 1, so the last term ($\ln 1 = 0$) in (3) disappears.

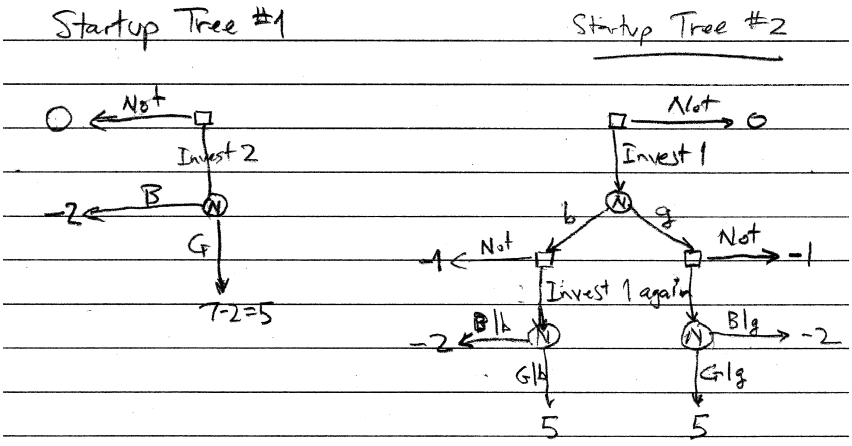


Figure 1: Your decision nodes are marked with small boxes, and random event nodes ("Nature" moves) by circles marked N.

- In Bayesian statistics, one explicitly estimates the posterior distribution given some prior distribution and given the likelihood functions for the observed data.

5. Behavioral economists find that in many situations, people tend to underweight prior information, i.e., over-respond to news. Did you fall into this trap on the zika example? The frequency intuition helps many people get it right.

III. Dynamic Decision Trees

A. Here is a pair of example problems to get us going.

Ex # 1: You decide now whether to invest 2 ($ million) in a startup company.

- With probability $p(A) = .25$ it will succeed and be bought out at a price whose present value is 7.
- With probability $p(B) = .75$ it will fail and have salvage value 0.
- See Startup Tree #1.

Ex # 2: You decide now whether to invest 1 ($ million) to allow the startup to release

a beta version of its product. After seeing whether the beta version is good or bad, you can decide whether to invest the remaining 1 ($ million).

- See Startup Tree #2.
- The likelihoods of a good beta version for a product ultimately fated for success and for failure are $p(a|A) = 0.9$ and $p(a|B) = 0.3$.
- Using Bayes Theorem etc, the spreadsheet calculates beta version probabilities $p(a) = 0.45$ for good, and $p(b) = 0.55$ for bad; and posterior probabilities $p(A|a) = 0.5, p(B|a) = 0.5, p(A|b) = \frac{1}{21}, p(B|b) = \frac{20}{21}$.

B. Now let's solve the two example problems.

- In Tree #1, if we invest then we get expected payoff $E\pi = \sum_{s\in S} p(s)\pi(s) = p(A)\pi(A) + p(B)\pi(B) = .25(7-2) + .75(-2) = -0.25$. This is worse that the 0 payoff to Not investing.
- Hence we choose the Not invest branch and get payoff 0. See solved Tree #1.
- To solve Tree #2, we begin by figuring out what we would get if we invested following the beta release.
- If the release is good, then we would get $E[\pi|a, inv] = \sum_{s\in S} p(s|a)\pi(s) = p(A|a)\pi(A) + p(B|a)\pi(B) = .5(7-2) + .5(-2) = 1.5$.
- If the release is bad, then we would get $E[\pi|b, inv] = \sum_{s\in S} p(s|b)\pi(s) = p(A|b)\pi(A) + p(B|b)\pi(B) = \frac{1}{21}(7-2) + \frac{20}{21}(-2) = -\frac{35}{21} = -\frac{5}{3}$.
- Comparing $E[\pi|a, inv] = 1.5$ to $E[\pi|a, Not] = -1$, we decide to invest if the beta release is good, resulting in expected payoff $E[\pi|a] = 1.5$.
- Comparing $E[\pi|b, inv] = -\frac{5}{3}$ to $E[\pi|a, Not] = -1$, we decide Not to invest if the beta release is bad, resulting in expected payoff $E[\pi|b] = -1$.
- Given this contingency plan, we see that the expected payoff to sinking an initial investment of 1 is $E[\pi|inv] = \sum_{m\in M} p(m)E[\pi|m] = p(a)E[\pi|a] + p(b)E[\pi|b] = .45(1.5) + .55(-1) = .125$.
- Since this expected payoff exceeds the 0 payoff for Not sinking the initial

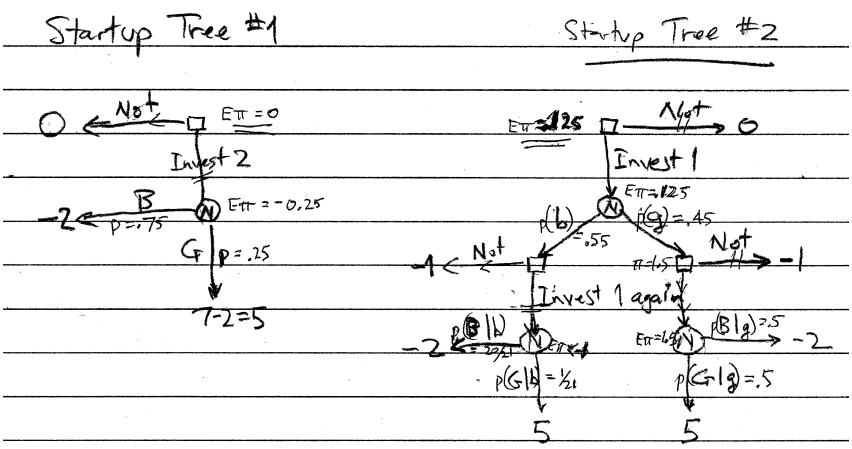


Figure 2: Solution takes $E\pi$ at N-nodes and maximizes at your decision nodes. Double slashes denote eliminated branches at each of your nodes.

investment, we see that our optimal course of action is: *sink the initial 1, and invest another 1 if the beta release is good, otherwise to abandon the startup.*

- See solved tree #2 in Figure 2.

IV. These examples illustrate the general rule for solving dynamic decision problems. Here is the cookbook.

A. First, construct the decision tree as follows.

1. Consider the logical sequence of events and the important contingencies, described as messages, states and choices. What will you know at the time you make each choice, and what are the possibilities after?
2. Sketch the resulting tree, carefully distinguishing your choice nodes from Nature nodes (where messages or states are revealed).
3. Fill in the probabilities of branches emanating from each N-node, using Bayes Theorem as necessary.
4. Fill in the final payoffs at terminal nodes (the most remote branch tips), contingent on all prior choices and events.

B. Then solve the decision tree, using Backward Induction as follows.

1. Solve each node just prior to terminal nodes.
   - If it is a N-node, then take the expected payoff over all branches.
   - If it is a C-node, then take the maximum payoff over all branches.
2. Replace that node and all its branches by a terminal node whose payoff is the solution just computed.
3. Iterate the previous steps on the truncated tree, keeping track of where each node solution came from. This is called induction.
4. When you get to the initial node you are done, and you know the expected payoff for the optimal plan. Write out that plan, i.e., the node-by-node solution to the decision tree.

C. Writing and solving trees fluently takes practice. Try a few practice problems and try your own examples.

D. So far we have represented the final payoffs as profits, and we will continue to do so for the sake of simplicity. But this implicitly assumes risk neutrality, and the offi-

cial theory says that the payoffs really are the utility of the final outcomes. [DF's personal opinion: In most practical applications, the extra layer of complication is not worth the trouble.]

E. The economic value of information is the expected payoff in the solved tree when the information is available minus the expected payoff in the solved tree for which the information is not available.

- You can think of Startup tree #2 as informed, and #1 as uninformed.
- In this case, the value of information is $V = 0.125 - 0.0 = 0.125$, or \$125,000.

F. Notice that information only has economic value to the extent that it leads you to alter a decision.

- In the Startup example, you would make the second stage investment following a good beta version, but without information you would not invest.
- In a different example where no message would alter your optimal plan, the economic value of information would be zero. (Of course, it might have some psychological value.)

V. Extensions of Decision Theory

A. The basic recipe is always the same: specify the sequence of information events and decisions to create a well-posed optimization problem, and solve the problem by backward induction.

B. Many decision problems involve trading off immediate payoffs against future payoffs.

1. Backward induction here leads to the following general insight.
2. When solving a decision node, consider the "continuation value" obtained from downstream (already solved) nodes.
3. The optimal choice at each node maximizes the sum [current payoff plus continuation value].

C. Bellman's equation is a widely used tool in discrete time optimization.

1. It is a recursion equation that basically says the value at time t is the maximum among current choices of [current payoff plus continuation value].
2. It can be written for infinite horizon (i.e., the last period is not known from the outset) problems where backward induction cannot be done directly.
3. A few of numerous examples include
   a. job search: at each stage, compare the payoff to accepting the best job currently on offer to the continuation value of continuing the search
   b. deferral options: compare the payoff to making an irreversible investment now to the continuation value, knowing that better information will arrive in the meantime (this has the same flavor as Startup tree #2)
   c. Financial call options: at each stage, compare the value of exercising the option (to purchase the underlying financial asset at the given strike price) to the continuation value of waiting (possibly enabling the underlying price to go higher and thus provide a higher net payoff).

D. A standard technique to evaluate numerically the value of financial options is to build a large but finite binomial event tree

- Each N-node is has 2 branches: the underlying asset price goes up or down by a small percentage (say .01) with known probability.
- Each C-node also has 2 branches: exercise the option or let it run.
- You solve the tree exactly as above.
- Example: MS thesis of Cagney and Oh, about 20 years ago.

7. Monopoly and Price Discrimination

See Varian Ch 14.

I. Simple Monopoly vs. Perfect Competition

A. The core behavioral rule is just the same...

The firm maximizes profit,

$\max py - c(y)$

B. The resulting decision rule is the same...

The firm still sets marginal revenue equal to marginal cost.

C. But the things the firm has control over have now changed... We no longer assume that the firm is a price taker.

1. Mathematically, this means that there is some functional relationship between the quantity the monopolist produces and the price the monopolist can charge.
2. The firm's problem is now subtly different:

$\max p(y)y - c(y)$

3. In particular, marginal revenue is not pegged to some exogenously given price.

II. The Monopolist's Problem

A. When a competitive firm decides to change output, she only has to think about one effect on revenue.

- Revenue increases by $p\ dy$

B. Monopolists have to trade off two effects:

1. Revenue *increases* by $p\ dy$.
2. In order to sell more units, the monopolist has to lower price.

- Revenue *decreases* by $ydp = y\frac{\partial p}{\partial y}dy$
- Unless the monopolist can price discriminate, this revenue change affects all of the monopolist's units – not just the marginal ones.
- These non-marginal units are called inframarginal.

3. Summing these effects we get :

$$dR = d(py) = pdy + ydp = [p + \frac{\partial p}{\partial y}y]dy \tag{1}$$

C. So Marginal revenue is

$$\begin{aligned} \frac{dR}{dy} &= [p + \frac{\partial p}{\partial y}y] \\ &= p[1 + \frac{\partial p}{\partial y}\frac{y}{p}] \\ &= p[1 + \frac{1}{\epsilon(y)}], \end{aligned} \tag{2}$$

where $\epsilon(y) = \frac{\partial y}{\partial p}\frac{p}{y}$ is own price demand elasticity.

III. Profit Maximization

A. The first order condition for the firm's problem sets marginal revenue equal marginal cost.

$$p(y)[1 + \frac{1}{\epsilon(y)}] = c'(y) \tag{3}$$

Ex: The linear case.

Ex: Const Elastic Demand

B. $|\epsilon(y)| > 1$ at the Monopolist's optimal production level y.

Otherwise we'd have negative MR, which implies that reducing production increases profit.

IV. Deadweight Loss

A. When firms are competitive, marginal costs of production equal (inverse) demand.

1. Marginal revenue is wherever the supply curve happens to intersect the demand curve.
2. This is a consequence of the price taking assumption.

B. If demand slopes downward, the monopolist's marginal revenue curve will lie below the demand curve.

$p(y) > p(y)[1 + \frac{1}{\epsilon(y)}]$

Since MC is upward sloping, we conclude that a monopolist will cut back output relative to a price-taker.

C. Competitive firms produce TS-max level and therefore monopoly power *must result in deadweight loss.*

D. Caveat: price discrimination can reverse this effect.

Ex: Quasilinear case.

V. Passing Along Costs (or Taxes)

A. What happens to prices if the costs of production change?

B. To make things simple, we assume linear costs.

1. This is a very common assumption in industrial organization.
2. In this case it makes it easy to substitute cost increases for tax increases.

C. In this case it can be shown that

$$\frac{\partial p}{\partial c} = \frac{1}{2 + yp''(y)/p'(y)} \tag{4}$$

D. So...

1. With linear demand, half of costs are passed on in the price.
2. With const elasticity demand, price goes up by a proportion *more than 1.0!* (see Varian, p.237.)

VI. Where Do Monopolies Come From?

A. Natural Monopolies

1. A natural monopoly is an industry with a cost structure that prohibits marginal cost pricing.
   a. Competitive production and pricing would lead to negative profits.
   b. Regulating price = MC would drive firm out of business.
2. Occurs when fixed costs are extremely large relative to marginal costs, or otherwise have IRS or decreasing AC up to very large scale.
3. Traditional regulatory solution is to force monopolists to price at average cost.
   a. Allow monopolists to just break even.
   b. Still inefficient.
   c. Gives regulated firm an incentive to inflate costs.

Ex: Natural Monopoly

B. Minimum Efficient Scale

1. What output level minimizes AC?
2. If there is enough demand to accommodate multiple firms producing at this cost, then an industry can be competitive.
   Otherwise only a single firm can survive.
3. Notice that MES is relative to technology

C. Unnatural Monopoly: Barriers to Entry

1. Explicit barriers
2. Raising rivals cost
3. Price controls
4. Product standards

D. Problem set asks: Which of these firms are natural or unnatural monopolists? What are reasonable public policies towards them? Google, Facebook, AirBnB, Amazon, Uber.

VII. Price Discrimination Overview

A. Price discrimination is the sale of identical units of a good at different prices.

B. By price discriminating, a firm can capture some of what would be consumer surplus. PS ↑.

1. In so doing, a monopolist may also increase output, leading to a more efficient outcome. So it is possible that we also have CS ↑.

C. Constraints on price discrimination

1. The firm must have market power; otherwise it will just be a price-taker.
2. Arbitrage must somehow be limited; otherwise low price units could be resold and undercut the higher priced units.
3. The firm must somehow be able to detect WTP differences across consumers, and/or across units purchased by a single customer.
4. There may also be legal or moral constraints.

D. The three classical types of price discrimination are methods of coping with the constraints and sorting consumers according to their WTP.

VIII. A Basic Model from Varian. (Not required to master it in detail.)

A. A simple quasilinear model helps explain several varieties of price discrimination.

B. Consumers, $i = 1, 2$ have utility $u_i(x) + y$, normalized so that $u_i(0) = 0$. .

1. Think of y as money left over for everything other than x.
2. Consumers are willing to pay up to $u_i(x)$ for x units of the good.
3. Hence i's marginal WTP is $u_i'(x)$.

C. The inverse demand curve for the individual consumer is therefore found by solving the consumer's problem

1. max $u_i(x) + y$
   s.t. $px + y = m$
2. FOC is $p = u_i'(x)$, the inverse demand curve.
3. In other words, i's marginal WTP is $u_i'(x)$.

D. From now on we'll assume that consumer 2 has higher WTP than consumer 1.

1. $u_2'(x) > u_1'(x)$, so by integration,
2. $u_2(x) > u_1(x)$.
3. This implies the **single crossing property**: any indifference curve for one consumer crosses an IC of the other consumer only once.

E. To avoid messy expressions, assume that the monopolist has constant marginal cost, so $c(x) = cx$.

IX. Perfect Price Discrimination (aka First Degree)

A. The monopolist is able to charge a different price for each unit sold.

B. To see the implications, suppose the monopolist makes an offer to each buyer i of x_i units for a lump sum payment of r_i.

$$\max_{r \geq 0} \quad r - cx \text{ s.t. } u_i(x) \geq r. \tag{5}$$

C. Constraint holds with equality at optimum.

D. The FOC is $u'(x^*) = c$ .... which is Pareto efficient!

E. So $r^* = u(x^*)$.

1. Note that this x^* is the same level of output as a competitive firm.

2. $u'(x) = p(x) = c$.

F. This lump sum solution is equivalent to charging a different price (for marginal willingness to pay) for each unit of the good.

G. Constraints: all of them are problematic here.

H. Colleges attempt to approximate this for families who apply for financial aid. First degree price discrimination is not necessarily an evil plot by the producer. Given high fixed costs, it may be the only way for the producer to stay in business.

X. Second Degree Price Discrimination

A. The monopolist charges prices that are not simple per-unit prices.

1. Sometimes called nonlinear pricing.
2. Includes quantity discounts, and block pricing as in local water bills. Also gold/silver bronze health plans or data plans.

B. Simplest version: a monopolist offers two different price/quantity *bundles* (r_i, x_i)

1. Bundle i is designed for consumers of type i.
2. The monopolist doesn't know whether a given consumer is type 1 or type 2.
3. The pricing scheme encourages consumers *sort themselves*.

C. The following analysis is *optional* for MS students (but PhD students are supposed to master more general versions.)

D. In order to get type i consumers to choose the targeted bundle i, the monopolist has to satisfy two types of constraints:

1. Individual Rationality (aka participation):

$u_1(x_1) - r_1 \geq 0$ ***
$u_2(x_2) - r_2 \geq 0$

2. Self selection (aka. incentive) constraints:

$u_1(x_1) - r_1 \geq u_1(x_2) - r_2$
$u_2(x_2) - r_2 \geq u_2(x_1) - r_1$ ***

1. A profit maximizing producer wants to set r_1 and r_2 as high as she can while satisfying the constraints.
2. This fact combined with the single crossing property guarantees that some of the constraints above bind, i.e., hold with $=$, not with $>$. It turns out (see Varian) that the binding constraints are the two marked ***. Thus we have

$r_1 = u_1(x_1)$
$r_2 = u_2(x_2) - u_2(x_1) + r_1 = u_2(x_2) - u_2(x_1) + u_1(x_1) < u_2(x_2).$

3. That is, charge the low value consumer his max WTP for the low target bundle, and charge the high value consumer the highest price that doesn't cause him to switch away from the high target bundle.

E. The monopolist's problem

1. The monopolist gets the sum of the profits from the two consumer types.

$\pi = r_1 - cx_1 + r_2 - cx_2$

2. All of our hard work above gives us constraints to substitute into this equation:

$\pi = u_1(x_1) - cx_1 + u_2(x_2) - u_2(x_1) + u_1(x_1) - cx_2$

3. We can maximize this with respect to outputs x_1 and x_2, to obtain the FOCs below.

F. Welfare

1. The FOC $0 = \frac{\partial \pi}{\partial x_1}$ yields

   $u_1'(x_1) = p(x_1) = c + u_2'(x_1) - u_1'(x_1)$

2. This tells us that the per unit price charged to the low value consumer is above marginal cost, implying implying a DWL.
3. The other first order condition, $0 = \frac{\partial \pi}{\partial x_2}$, yields

   $u_2'(x_2) = p(x_1) = c.$

4. This tells us that the per unit price charged to the high value consumer is equal to marginal cost. No efficiency loss here.

G. Conclusion: To max profit, target a bundle to high value consumer (#2) s.t. price (on last unit) = MC. Find a bundle for low value consumer (#1) that cuts back from the efficient quantity, is (barely) not preferred by #2, but exhausts #1's WTP.

H. Remark. I am unaware of any firm that actually does such calculations to obtain their price/quantity menu. It's hard to estimate the preferences u_i, and estimation errors could throw the calculation way off. Yet this approach gives insight to the menu that firms may settle upon after trial and error.

XI. Third Degree Price Discrimination (aka Market Segmentation).

A. This case is the most realistic, and the calculations parallel the way some firms actually think about it. You should be able to master this case.

B. The monopolist is able to charge different prices to identifiably different groups but not able to charge different prices within any group.

1. Think student discounts, or senior citizen discounts.
2. Or last minute shoppers, or domestic vs foreign market.

C. Assume for now that arbitrage is not possible.

1. The monopolist's problem is

   $\max p_1(x_1)x_1 - cx_1 + p_2(x_2)x_2 - cx_2$

2. The FOCs from this problem can be written as:

   $p_1(x_1)[1 - \frac{1}{|\epsilon_1|}] = c$
   $p_2(x_2)[1 - \frac{1}{|\epsilon_2|}] = c$

3. So we can write $p_i = M_i c$ where markup factor is $M_i = \frac{1}{1-\frac{1}{|\epsilon_i|}} = \frac{|\epsilon_i|}{|\epsilon_i|-1}$.
4. In particular, $p_1(x_1) > p_2(x_2)$ only when $|\epsilon_1| < |\epsilon_2|$
5. What if $|\epsilon_1| < |\epsilon_2|$ but arbitrage is possible, at unit cost k?
   a. If the formulas above give $p_1 \leq p_2 + k$, then arbitrage is unprofitable and has no impact.
   b. But if they give $p_1 \geq p_2 + k$, then profitable arbitrage will undermine that form of price discrimination.
   c. In that case, the profit-maximizing choice can be found by writing $p_2 = p, p_1 = p + k$, putting this into a profit function for the firm, and finding the profit-maximizing p.

D. Welfare

1. Does the ability to price discriminate in the third degree help or hurt social welfare (TS)?
2. This depends on the effects on output. Varian shows that it can go either way (or be neutral).
   a. First, the only way welfare can be *improved* is if output increases due to the discrimination.
   b. Second, if prices and output changes (relative to ordinary monopoly) satisfy $(p_1 - c)\Delta x_1 + (p_2 - c)\Delta x_2 > 0$, then welfare has to improve!
   c. Third, if a whole new market is served due to the discrimination, welfare has to improve.

XII. Other forms of price discrimination

- 2-part tariffs
  A. membership fee plus (low) per unit price.
  B. if all customers are identical, can obtain same result as perfect price discrimination by charging $p = c$ and fee=CS.
  C. behavioral economics angle: gym membership.
- bundling
  A. season tickets, package tours, ...
  B. can increase firm's profits to the extent that consumers differ in their WTP for bundle components.
  C. in that sense, covers the opposite case from 2-part tariffs.
- loyalty programs
  A. usually take the form of quantity discounts, with a time lag
  B. frequent flyer miles are partially convertible to currency
  C. generally try to blunt direct price competition
- peak-load pricing
  A. Uber is notorious for using this sort of price discrimination.
  B. It can be thought of as market segmentation, but
  C. there may also be differences on the cost side.
  D. an extreme recent version: texas utility makes electricity free after 9pm (NYT NOV. 8, 2015).

8. Oligopoly

Some of this material can be found in Varian Chapter 15-16; see especially 16.5-16.10.

I. Overview

A. So far we have only looked at two extreme types of markets.

1. Competitive markets have only price-taking firms (presumably lots of them).
2. Monopolist markets have one firm with unilateral pricing power.

B. We now look at markets with firms that have some pricing power, but not unilateral.

Oligopoly: from Greek, more than one but less than "many."

C. We use game theory to study behavior in oligopolies.

1. Firm decisions affect one another $\iff$ strategic interaction.
   a. Need an equilibrium concept that describes multiple agents trying to optimize.
2. Game theory yields many surprising conclusions. Here's the first.
   a. We solved the monopolist's problem by describing its choice of quantities.
   b. We could have just as easily (and with the same result) had the monopolist choose a price.
   c. This symmetry disappears in our study of oligopoly: the equilibria turn out to be quite different.

II. (Normal Form) Game theory

A. A normal form (simultaneous play) game (NFG) is defined by three elements

1. A list of N players
2. A set of strategies for each player, $s_i \in S_i$. E.g.,$S_i = \{s_{i1}, s_{i2}, ..., s_{ik}\}$.

3. A payoff function for each player, $\pi_i(\mathbf{s})$, where the profile of all players' strategies is $\mathbf{s} = (s_i, \mathbf{s_{-i}})$.

B. Let $\mathbf{s_{-i}}$ be a vector of strategies of all players other than i. The **best response function** (or correspondence) is $BR_i(\mathbf{s_{-i}}) = argmax_{s_i \in S_i} \pi_i(s_i, \mathbf{s_{-i}})$. In words, for a given profile $\mathbf{s_{-i}}$ of other players' strategies, player i's best response is the strategy (or subset of strategies) that maximizes his payoff.

C. A **Nash equilibrium** is a strategy profile $\mathbf{s}^*$ in which every player is making a best response to the other players' strategies, i.e.,

$$s_i^* \in BR_i(\mathbf{s_{-i}^*}), \quad i = 1, ..., n. \tag{1}$$

.

D. These definitions are quite general and apply in politics, biology, business, traffic engineering, etc. etc. Here we will apply them to oligopoly.

III. Quantity Setting: Cournot Markets

A. The Duopoly NFG

1. $N = 2$ players, called firms.
2. Strategy is the output quantity $y_i \in [0, \infty) = S_i$.
3. The choices y_1, y_2 are made simultaneously (logically speaking).
4. We'll keep things simple in computing the payoff functions (profit functions). Set $Y = \sum_{i=1}^{N} y_i$ to be total output, and assume a linear inverse demand curve (for perfect substitutes)

   $p = a - bY$

5. We'll also assume a linear cost curve, i.e., identical marginal cost c for all firms and zero fixed cost.

6. Then the profit to any firm i is:

$$\pi_i = y_i(a - bY) - y_i c = (a - c - bY)y_i. \tag{2}$$

7. N.B. Aggregate output quantity of other firms $Y_{-i} = Y - y_i$ affects firm $i's$ profits and therefore his optimal choice.

B. Best Response Function

1. Here the best response function BR_i describes firm i's best choice of quantity y_i as a function of the quantity choices of everyone else.
   a. Note this doesn't imply that firms actually *know* the quantity choice of others.
   b. Action is simultaneous.
2. We will see that in this quantity setting game, $\frac{\partial BR_i}{\partial y_j} < 0$
   a. If you know your rival's output is high, you want your output to be low.
   b. Quantity is a *strategic substitute*.

   Ex: The FOC for (2) is

$$0 = \frac{\partial \pi_i}{\partial y_i} = a - c - 2by_i - bY_{-i}. \tag{3}$$

   Solving for y_i, we get the BR function

$$BR_i(Y_{-i}) = [\frac{a-c}{2b} - \frac{1}{2}Y_{-i}]_+ \tag{4}$$

   where $[z]_+ = \max\{0, z\}$. See Figure 1.

C. Nash Equilibrium

1. A Nash equilibrium is a profile of strategies at which no player has an incentive to change their behavior given what others are doing.
2. Or, all firms are simultaneously playing their best response.

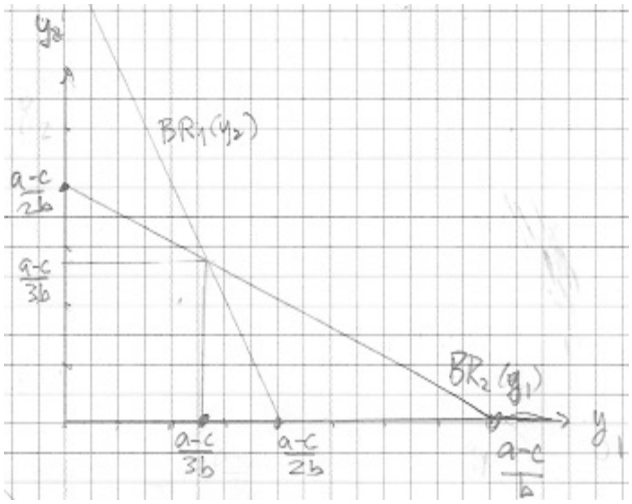


Figure 1: BRs and NE in Cournot duopoly.

3. Adding the FOCs (3) for $i = 1, ..., m$, we get

   $0 = N(a - c) - 2bY - b(N - 1)Y$,

   so total output in Cournot Nash equilibrium is

   $Y^* = \frac{N}{(N+1)b}(a - c)$, and by symmetry,

   $y_i^* = \frac{Y^*}{N} = \frac{a-c}{(N+1)b}$.

   For duopoly, $y_i^* = \frac{a-c}{3b}$, and NE payoff is

   $\pi_i^* = (a - c - bY^*)y_i^* = (by_i^*)y_i^* = \frac{(a-c)^2}{9b}$.

D. Asymptotics

1. Can we describe the equilibrium behavior of Cournot firms in terms of the number of firms (N)?

   a. By doing this we can look at the relationship between oligopolies and both monopolies and competition.

2. Denoting s_i as $\frac{y_i}{Y}$, profit maximization gives us Marginal Revenue = Marginal Cost. Using familiar tricks on Marginal Revenue, we get

   $p(Y)(1 + \frac{s_i}{\epsilon}) = c_i'(y_i)$

3. If all firms have the same constant marginal cost c and fixed costs that they can cover in Nash equilibrium, then $s_i = 1/N$ and

   $p(Y)[1 + \frac{1}{N\epsilon}] = c.$

4. The main result is a price somewhere between competition and monopoly:
   a. If $N = 1$ this price is just the monopoly price.
   b. As N approaches infinity, price converges to the competitive level.
   c. With N in between, price remains above marginal cost, but below the monopoly level.

E. A problem with Cournot Analysis

1. We usually think of firms setting price – not quantity.
2. Where do prices come from in Cournot markets?
   a. They come from the inverse demand curve, but what does that mean?
   b. The Cournot model implicitly assumes that firms just dump their output on the market and accept the market clearing price.
3. Bertrand (1883) criticized that assumption of the Cournot (1838) model, and advised assuming directly that firms set prices. As we will now see, the predicted (equilibrium) outcome is quite different.

IV. Price Setting: Bertrand Markets

A. Describing a Duopoly

1. Two firms simultaneously choose price, given a demand curve $D(p)$.
2. Assume constant marginal costs c_i, possibly different across firms.
3. Consumers will buy from the lower priced firm, since they produce perfect substitutes.

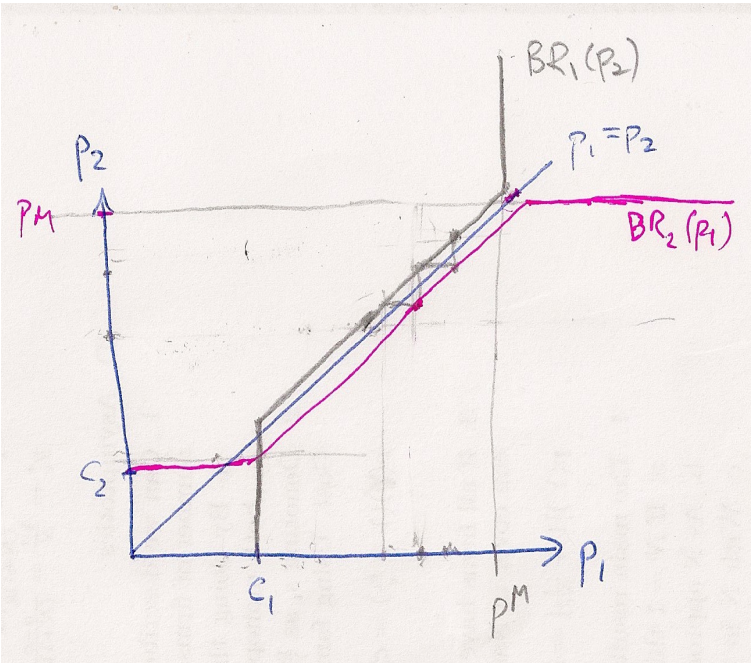


Figure 2: NE in Bertrand Model. Monopoly price is p_M, firms have marginal costs c_1, c_2.

4. Firm i's demand is given by

   a. $D(p_i)$ if $p_i < p_j$

   b. $D(p_i)/2$ if $p_i = p_j$

   c. 0 if $p_i > p_j$

B. Best Response

1. That discontinuity in demand leads to a discontinuity in payoff (profit) functions. Taking FOC's won't work well here.

2. The BR is to undercut the rival firm's price very slightly, but never price above the monopoly price or below own MC. See Figure 2.

3. Note now that the strategic variables – p_i – are *strategic complements*: the lower a rival's price, the lower you'd like your price to be.

4. By contrast, Figure 1 shows that in Cournot duopoly, the lower your rival's output, the *higher* is your best response output — a case of *strategic substitutes*.

C. Nash Equilibrium

1. A Nash Equilibrium in the Bertrand game is a set of prices at which no firm has an incentive to change his or her price.
2. Assume, without loss of generality, that $c_j > c_i$.
3. Then in a Nash equilibrium:
   a. $p_i = c_j$ (actually, a tiny bit lower)
   b. $p_j \geq c_j$
4. The price in the market is competitive – it is equal to (the second lowest) marginal cost.
   a. This is especially striking if we assume (as we often do) that firms share a common marginal cost.
   b. If $c_i = c_j = c$ then we have $p_i = c$.
5. In the price setting game, then, the prediction is the competitive outcome *even with only 2 firms*.

D. The Bertrand Paradox

1. Some call that last result the "Bertrand paradox."
2. Intuition tells us that, say, five firms should compete much more fiercely than two firms.
   a. Indeed there is experimental evidence to this effect.
3. One way of to get less competitive behavior in pricing games with few firms is to feature repeated games.
4. Another is to have firms selling slightly different products, as we now consider.

V. Differentiated goods variants of Cournot and Bertrand

A. More realistically, the goods are not *perfect* substitutes.

B. The Cournot variant writes the inverse demand functions for two goods like this:

$$p_1 = A_1 - B_1 y_1 - C y_2 \quad (5)$$

$$p_2 = A_2 - C y_1 - B_2 y_2 \quad (6)$$

- The cross price effect C must be the same for both firms by an obscure aspect of demand theory.
- Imperfect substitutes if $C > 0$, complements if $C < 0$, independent if $C = 0$.
- The special case $A_1 = A_2, B_1 = B_2 = C > 0$ reverts to perfect substitutes.
- The special case $C = 0$ reverts to two monopolies (in unrelated goods).

C. For constant marginal cost c for everyone, the usual optimization problem for BR has FOC that yields $y_1^* = \frac{A_1 - c - C y_2}{2B_1}$, and similarly $y_2^* = \frac{A_2 - c - C y_1}{2B_2}$.

D. Comparing the usual case $0 < C < B_1, B_2$ to perfect substitutes, you can see that the BR lines rotate outward, and NE outputs and profits increase, as substitutability C decreases.

E. Moral of story: firms prefer to differentiate their products.

F. There is a parallel analysis for Bertrand, using the direct demand functions. It is still true that prices are strategic complements while outputs are strategic substitutes.

G. These variants are more useful in applied work than the original pure substitute models.

VI. (Extensive Form) Game Theory: Basic ideas, assuming perfect info.

VII. Quantity Setting With A Leader: Stackelberg Markets.

- Draw "tree" for 2 player sequential game, each with three possible output levels. Solve by backward induction.

- Draw tree for 2 player sequential game, each with an interval $[0, p^M]$ of possible output levels. Solve by backward induction: last mover chooses BR. Previous mover predicts this, and optimizes.
- Take the duopoly example with linear demand, const MC, say $N = 2, a = 30, b = 1, c = 6$.
- For comparison, Cournot-Nash equilibrium is $y_1^* = y_2^* = \frac{30-6}{3} = 8$, price is $p^* = a - bY = 30 - 16 = 14$, and profits are $\pi_1^* = \pi_2^* = (p^* - c)y_i^* = 8 * 8 = 64$.
- The Stackelberg leader, firm 1, chooses output y_1 to maximize her profit, knowing how the follower will react. That is, she assumes that $y_2 = BR_2(y_1)$.
- Using equation (4), we see that $BR_2(y_1) = \frac{a-c}{2b} - \frac{1}{2}y_1 = 12 - \frac{1}{2}y_1$.
- Hence she solves

$$\begin{aligned} \max_{y_1 \geq 0} \pi_1(y_1, BR_2(y_1)) &= (a - c - b(y_1 + BR_2(y_1)))y_1 = (30 - 6 - (y_1 + 12 - \frac{1}{2}y_1))y_1 \\ &= (12 - \frac{1}{2}y_1)y_1, \end{aligned} \tag{7}$$

- which is easily seen to have solution $y_1^{SB} = 12$.
- Hence $y_2^{SB} = BR_2(y_1^{SB}) = 12 - \frac{1}{2}y_1^{SB} = 6$, and $p = 30 - (12 + 6) = 12$, so $\pi_1^{SB} = (12 - c)12 = 72$ and $\pi_1^{SB} = (12 - c)6 = 36$.
- Compared to Cournot NE: price is lower, profit for leader is higher, but follower profit and total profit are lower in Stackelberg NE.

VIII. Collusion and cartels

- If all the firms in an industry could agree on the total output level Y, what would they choose?
- This is just a restatement of the basic monopoly problem.
- But there are difficulties in implementation.

  - Firms might disagree on how to split up Y into quotas.
  - Even if they agree in principle, it is true in practice that they all have an incentive to exceed the quota – the firm producing the excess gets all the benefit (pdy) but absorbs only a fraction of the cost Ydp in MR.
  - This is true whether or not the other firms stick to their quotas, as long as Y is less than in Cournot equilibrium.

- To enforce quotas, cartels need to reliably detect violations and punish them.
- They also need to prevent new entry.
- Theory of repeated games (not included in this course) says that a key is the discount factor, based on likelihood that the market will continue, and interest rates.
- Of course, cartels are discouraged by anti-trust laws, at least in major industrialized economies during the last century.
- Historically speaking, most successful cartels needed strong government support.

IX. Kinked Demand Curve and Sticky Prices

- Paul Sweezy (1939) argued that most oligopolies work differently than in previous models.
- In industries ranging from autos to zinc, there was an established price $\bar{p}$ that seldom moved.
- Firms anticipated that if they cut price below $\bar{p}$, their rivals would match them (as in the steeper part of the demand curve in Figure 3). But if they raised their price, nobody would follow them, so they would lose share rapidly (as in the flatter part of the demand curve in the Figure).
- Consequently, MR has a discontinuous drop at $D(\bar{p})$.

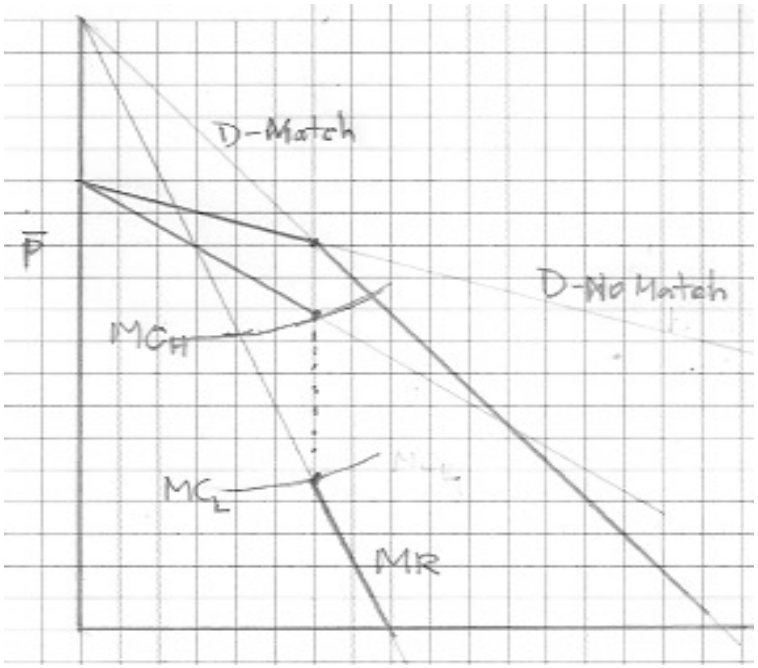


Figure 3: Demand and MR in KDC model.

- Profit-maximization therefore entails not changing quantity as long as MC lies between MC_L and MC_H in Figure 3). That is, prices are "sticky" — not responsive to moderate cost shocks.
- The theory is less popular today, partly because fewer industries now are like that, and partly because Sweezy's followers never explained where $\bar{p}$ comes from.
- My early paper "Producers' Markets: A Model of Oligopoly with Sales Costs," *Journal Economic Behavior and Organization*, 11:3 (May 1989) 381-398, offers an explanation of $\bar{p}$. Given its cost function, the battle for market share (via ads and other cost of sales), and the rational assumption of price matching on price decreases but not increases, the model shows that each each firm has a profit-maximizing value of $\bar{p}$. The lowest one among existing firms is the equilibrium value of $\bar{p}$, and that firm is the price leader.

X. Conjectural Variations.

Recall that, for homogeneous goods with inverse demand $p(Y) = p(y_1 + y_{-1})$, firm 1's

problem can be written as $\max_{[y_1 \geq 0]} y_1 p(y_1 + y_{-1}) - c(y_1)$. The first order condition fully written out is

$$\begin{aligned} c'(y_1) &= p(Y) + y_1 p'(Y)[\frac{dy_1}{dy_1} + \frac{dy_{-1}}{dy_1}] \quad (8) \\ &= p(Y) + y_1 p'(Y)[1 + \nu], \quad (9) \end{aligned}$$

where $\nu = \frac{dy_{-1}}{dy_1}$ is firm 1's *conjectural variation* — her belief about how a change in her output y_1 will affect the total output y_{-1} of all rivals.

- $\nu = 0$ is the Cournot conjectural variation. She takes as given her rivals' output level, and (incorrectly!) assumes that she can't affect it.
- $\nu = -0.5$ is the Stackelberg leader's conjectural variation in the simple linear duopoly. More generally, it can be the slope of other firms' summed reaction functions.
- $\nu = -1.0$ is the competitive or Bertrand conjectural variation. It ensures that price = MC, and says that the firm believes that other firms will replace any units it withholds from the market.
- $\nu = y_{-1}/y_1$ is the collusion conjectural variation – other firms will maintain their current share.
- "consistent" conjectural variations equate ν to the actual comparative statics of the model for each firm.

Economic theorists no longer find it fashionable to write down arbitrary expressions for ν, and the idea of consistent conjectures never got much empirical support. (But McGinty, 2016, may give it new life in the context of greenhouse gas abatement treaties.) Masters students now might find CVs useful in keeping track of various models of imperfect competition.

XI. Spatial Competition: Hotelling location models.

Let us now take a deeper look at imperfect substitutes. So far, we have taken as given substitution elasticities in utility functions and demand functions. We also noted (from the differentiated good Bertrand model) that firms tend to be more profitable when their products are less substitutable. How can we model that dimension of competition?

Hotelling (1929) apparently was the first to take up that challenge. His "Main Street" model used a spatial metaphor to describe the substitutability among products. Think of the producers choosing the products' characteristics (e.g., fuel economy, acceleration, and seating capacity of cars) in order to fill niches of the market that are relatively undersupplied. That is, firms choose location in the space of characteristics.

Hotelling considered a very special characteristic space: location within the interval $[0, 1]$. This could be taken literally as an address on a small town's Main Street, or metaphorically as in characteristic space.

We begin with a duopoly where firms choose location but not price; for simplicity we assume that price is fixed at $p_1 = p_2 = p > c$, where $c = c_1 = c_2$ is the constant marginal cost faced by both firms.

- Label the firms so that the location choices satisfy $z_1 \leq z_2 \in [0, 1]$.
- For simplicity, assume that consumers' preferred locations are uniformly distributed along $[0, 1]$.
- Also, for simplicity, assume linear transportation (or transformation) cost $t > 0$. Thus the delivered price at location z for firm j is $p_j(z) = p + t|z - z_j|$. The assumption is that consumers buy at the lowest delivered price.
- Under current simplifications, this means that firm 1 gets all customers in $[0, \hat{z})$ and firm 2 gets those in $(\hat{z}, 1]$, where $\hat{z}$ solves $p_1(z) = p_2(z)$. In other words, the market shares are $\hat{z}$ and $1 - \hat{z}$, where the customer at $\hat{z} = 0.5(z_1 + z_2)$ faces the same delivered price from both firms.

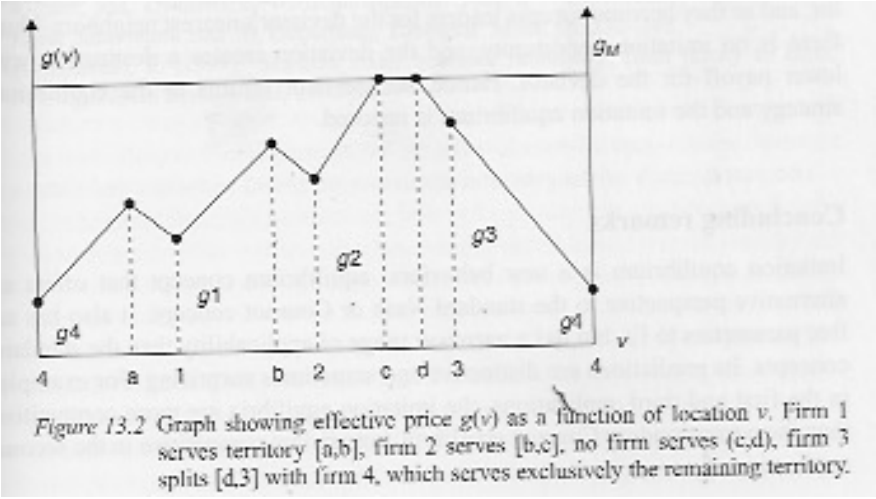


Figure 13.2 Graph showing effective price $g(v)$ as a function of location v. Firm 1 serves territory [a,b], firm 2 serves [b,c], no firm serves (c,d), firm 3 splits [d,3] with firm 4, which serves exclusively the remaining territory.

Figure 4: From Selten chapter in Friedman and Cassar (2004). Here $gi = p_i(z) - c$, and i is location z_i (denoted v in the Figure).

- Since $p > c$, firms maximize profit by maximizing market share.
- What is firm i's BR to location choice z_j of firm j? If $z_j < 0.5$, it is to locate a tiny bit to the right, at $z_j + \epsilon$. If $z_j > 0.5$, it is to locate a tiny bit to the left, at $z_j - \epsilon$. This is how i maximizes market share.
- So the unique NE of this Hotelling location game is for both firms to locate back-to-back at $z = 0.5$.

This simple game is sometimes used to explain (in part) why firms in a similar line of business tend to locate next door to eachother, and why political parties used to adopt very similar platforms.

There are many, many extensions of the model. Expanding the duopoly location problem above to triopoly yields no NE in pure strategies. With 4 firms, the pure strategy NE all have 2 firms back to back at $z = .25$ and the other 2 at $z = .75$.

What if firms first commit to specific locations and then pick prices? Figure 4 illustrates how the shares are determined from an arbitrary set of locations and prices for 4 firms. The analysis is a bit tricky because the endpoints of the line segment play a special

role. To make the location space more homogeneous (the technical word is 'isotropic'), one can join the endpoints to make a circle, and this is assumed in the Figure. It can be shown in this case that the unique NE in pure strategies is for firms to space themselves equally around the circle (maximum differentiation) and to all charge the price $p_i = c + t$.

Other variants of the Hotelling location game consider two dimensional locations, on a rectangle or (to make it isotropic) a torus and various numbers of firms. One can also consider nonlinear transportation (or transformation) costs. There seems to be room for applied work here, but I've not seen much published recently. There are ongoing laboratory experiments by UCSC PhD Curtis Kephart.

XII. Monopolistic Competition

Typically in 1-d space each product has two direct competitors, and perhaps lots of indirect competitors. In 2-d, there are typically at least 3 direct competitors, and often 5 or 6. In higher dimensions, i.e., if many characteristics are relevant, then there could be so many direct competitors it is not worth keeping track of them individually.

How to model competition against large numbers of direct competitors whose products are not especially close substitutes? It sounds complicated, but there actually is a pretty simple way to do it. Invented before game theory by Edwin Chamberlin and (apparently independently by) Joan Robinson in the early 1930's, it combines the free-entry assumption of perfect competition with markup pricing as in monopoly. Here's how the undergrad textbook version of Monopolistic Competition goes.

- Consider an industry, like restaurants or breakfast cereal, where each product is unique but has lots of not-very-close substitutes.
- Thus each firm faces a fairly elastic, but not infinitely elastic, demand curve.

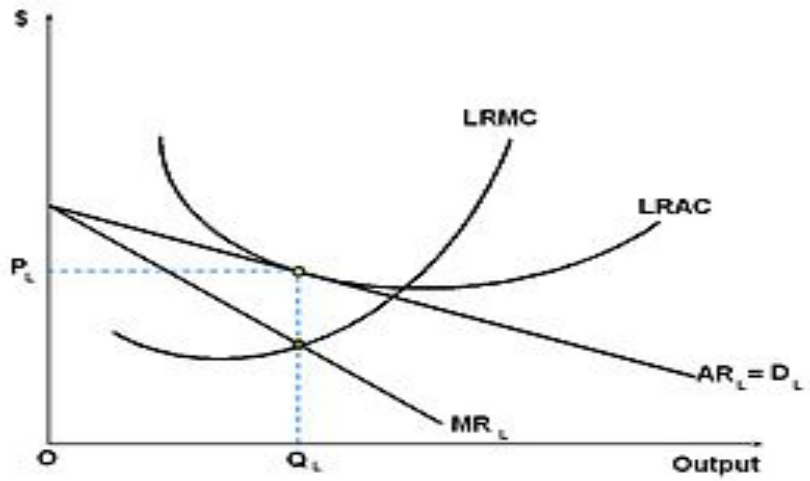


Figure 5: LR equilibrium for a Monopolistic Competitive firm. From Wikpedia article (not Wikip's finest).

- Solve the standard monopoly problem: the profit maximizing quantity is where MR=MC, and price is the inverse demand at that quantity.
- If economic profits are positive at that point, then more firms will be attracted into the industry, offering more fairly close substitutes. This will lower and flatten the demand curve, and lower the firm's maximum profit.
- Similarly, if profits are negative, then firms with similar products will exit, raising and steeping the demand curve, thus increasing its profit.
- In LR equilibrium, economic profit will be zero, as in Figure 5, since LRAC is equal to LRAR.
- Note that actually LRAC is tangent to inverse demand, aka LRAR. Since D is downward sloping, the firm is operating at less than efficient scale. In that sense, there is over-entry and inefficiency in monopolistically competitive industries.

There are also a grad level versions of monopolistic competition, which try to be more formal about the number of firms in the industry and where the demand curves come from. I can't find a really good version on the web, but you could start with

http://www.foltyn.net/wp-content/uploads/2009/12/dixitstiglitz.pdf .Here is my brief sketch, which might come in handy in your Macro class next quarter.

- Assume quasi-linear preferences for n goods, separable in good 0 (money): $u(x_0, x_1, ..., x_n) = x_o + G(f(x_1) + ... + f(x_n))$, where G and f are concave/convex functions.
- A typical choice is is the CES function $u = x_0 + [x_1^\rho + ... + x_n^\rho]^{1/\rho}$. These are sometimes called Dixit-Stiglitz preferences, after their 1977 originators.
- Using what you learned earlier this quarter, normalize price so $p_0 = 1$, and maximize u subject to a budget constraint. By quasilinearity, we can write the first argument as initial cash less expenditure. The resulting FOCs give the inverse demand equations

$$p_j = G'(\bar{x})f'(x_j), j = 1, ..., n, \tag{10}$$

  where $\bar{x} = f(x_1) + ... + f(x_n)$.
- This is nice, but for any finite, or even countable, number n of products, there is strategic interdependence, and so we have oligopoly, say a particular sort of Cournot oligopoly.
- To get true monopolistic competition, we have to take the continuum limit, so now $j \in [0, 1]$, and $\bar{x} = \int_0^1 f(x_j)dj$. The point is that (10) still works, so we can write direct demand as

$$x_j(p_j, \bar{x}) = \psi(p_j/G'(\bar{x})), j = 1, ..., n, \tag{11}$$

  where $\psi = f'^{-1}$ is the inverse of f'.
- That is, in the limit we have pricing power, with a monopoly markup independent of choices of other individual firms, but depending on an aggregate production index $\bar{x}$.

9. Public Goods and Externalities

See Varian Ch 23-24, and other micro textbooks for a more complete presentation.

I. Introduction

A. The invisible hand can go astray (i.e., CE may not be efficient) when activities of buyers and sellers affect ("external") third parties.

B. We'll consider general production and consumption externalities, and important particular sorts, such as public goods and network effects.

C. In principle, there are ways to restore efficiency in all these cases. In practice, it may be difficult or impossible.

D. Once you understand the principles, you may see opportunities for mutual gain in novel situations.

II. Public Goods

A. Public goods have one or both of the following special properties:

- Non-rival consumption: the amount I consume does not affect the amount potentially available for others to consume.
- Non-excludable consumption: once the good exists, it can be consumed by anyone (nearby), whether or not they pay for it.

B. Example. A local sports contest is held in a stadium with 50,000 seats, and is broadcast on live TV. The TV broadcast is both non-rival and non-excludable. If the signal were scrambled and viewable only by those who had a decoder box and paid for the key, then it would be excludable. The seats themselves are excludable and rival, so they are not public goods.

C. Other standard examples of public goods are national defense, public parks, and communication and transportation infrastructure. Infrastructure is congestible, so only partly non-rival.

D. The efficient quantity of a (nonrival) public good maximizes the *sum* of WTP's less production cost. Assuming continuous divisibility (e.g., air quality, or bandwidth), the FOC for efficiency is that the sum of WTP's for the last unit equals its marginal cost.

E. The CE amount of a non-excludable public goods is typically inefficiently low, since consumers can "free-ride."

III. A calculus example.

- Measure the quantity of the public good by its total cost Y, so $C(Y) = Y$.
- Agents $i = 1, ..., n$ have quasilinear utility $u_i(m, Y) = m + g_i(Y) = m + b_i \ln Y$.
- Thus each has WTP (or inverse demand) $g'(Y) = b_i/Y$.
- With non-rival consumption, the total social benefit is $B(Y) = b_T \ln Y$, where $b_T = \sum_{i=1}^{n} b_i$.
- Thus the efficient amount Y^o maximizes $B(Y) - C(Y) = b_T \ln Y - Y$. The FOC (which here is necessary and sufficient) is $0 = B'(Y) - 1 = b_T/Y - 1$, so $Y^o = b_T = \sum_{i=1}^{n} b_i$.
- A competitive market will provide the amount Y^c demanded by the most eager agent, and others will free ride. In this example, $Y^c = b_M$, where $b_M = \max_{i=1,...n} b_i$.
- For example, if there are $n = 100$ identical agents with $b_i = 5$, then $Y^c = 5$. Free riding causes a 99% shortfall from the social optimum $Y^o = 500$.

IV. Possible solutions to under-provision / free riding:

- If producers of the public good can make it excludable and can discover consumers' WTP, then they can restore efficiency by charging each agent the unit price $p_i = g_i'(Y^o)$, her marginal WTP. This is called Lindahl pricing.
- To spell it out, at price p_i, agent i will demand the efficient amount Y^o of the public good. In the previous example, $p_i = b_i/Y^o$, so she pays $p_i Y^o = b_i$. The

total amount purchased is $\sum_{i=1}^{n} b_i = b_T = Y^o$ — the efficient level of the public good !

- VCG mechanisms (discussed below) can elicit true WTP.
- In tight-knit communities, roughly efficient contribution levels for public goods can be attained via norms of behavior enforced by peer pressure.
- In mass societies, taxation can approximate efficient provision.
  - Suppose that agents vote on the per-capita tax, $s = Y/n$, and that the vote determines the outcome $Y = ns$.
  - Agent i's desired tax s_i solves $\max_x g_i(nx) - x$, so in the log example $s_i = b_i$.
  - In a simple voting model, the median desired rate $s_m = b_m$ will prevail.
  - In this case, $G^v = nb_m \approx \sum_{i=1}^{n} b_i = b_T$. The approximation is exact if median=mean, e.g., if the b_i distribution is symmetric.

V. Other sorts of externalities. General (albeit superficial) classification of externalities include

- positive (third parties benefit) vs negative (are harmed);
- consumption side (as in public goods) vs production side (spillovers from creating goods); and
- pecuniary (3rd party benefit or harm is due to changed prices) vs non-pecuiary (due to direct impact).

Here are some examples for you to classify in all three dimensions.

A. Increased UCSC enrollments drive up rents for non-student residents of Santa Cruz.

B. Driving your car increases the level of CO2 in the atmosphere.

C. Joining Facebook increases its value to your friends.

D. Drummers on West Cliff Drive affect your enjoyment of walking at sunset.

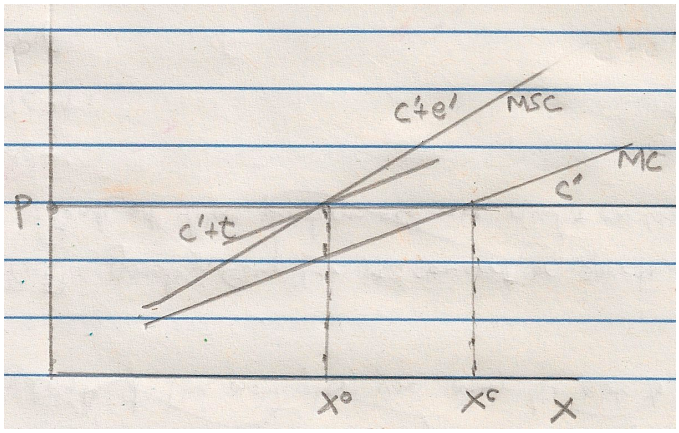


Figure 1: A negative production externality. Price-taking firm maximizes profit at x^c, but output level x^o maximizes total surplus.

E. Bee keepers increase the productivity of neighboring farms.

What can we say about the efficiency impact and remedies? Let's begin with a non-pecuniary negative production externality, such as a chemical plant polluting a nearby stream.

- The firm seeks to $\max_{x \geq 0}[px - c(x)]$, but output level x also creates costs $e(x)$ borne by downstream agents.
- Firm rationally chooses x^c that satisfies FOC $p = c'(x)$.
- Socially efficient choice x^o maximizes total surplus, $TS = px - c(x) - e(x)$. It typically satisfies the FOC $p = c'(x) + e'(x)$, i.e., price = marginal social cost, which includes the marginal external cost $e'(x)$ as well as marginal private cost $c'(x)$.
- Under the usual assumptions regarding costs ($c', e' > 0$ and $c'', e'' > 0$), we have $x^o < x^c$. That is, the competitive output (and pollution) level is higher than the socially optimal level, as in Figure 1.

How to remedy inefficiencies arising from externalities? The general idea is to somehow internalize the externality, so the invisible hand can resume its magic.

A. One way to internalize is to merge. The polluting firm could buy up all the downstream agents, and then it would recognize the entire cost $c(x) + e(x)$. Profit maximization now would be efficient. Of course, such mergers are often impractical.

B. Another perspective is to say that there is a missing market. If the polluting firm needed a permit to produce, and permits were sold in a competitive market, then (it can be argued) the market clearing price of the permit would be $e'(x)$. Then the firm's private MC would include $e'(x)$, restoring efficiency.

   - Should the missing market start by giving the polluters the right to pollute, or start by giving the downstream agents the right to clean water? Coase argued that it doesn't matter; whatever the assignment of property rights, the competitive outcome will be efficient. (His argument assumes quasi-linear preferences.)
   - Can you get a competitive market with only one agent (the polluter) on one side? Coase again argued that it doesn't matter; a bargaining process could also produce the same outcome.
   - Ronald Coase (1910-2013) won the 1991 Nobel prize, largely for the insight that, in principle, externalities can be internalized if property rights are properly established and transactions costs (in markets or bargaining) are negligible. The insight is very useful even though, as a practical matter, transactions costs are often prohibitive and the necessary property rights are unclear and/or unenforceable.

C. In some cases, the most practical way to internalize the externality is via a Pigovian tax (after British economist A.C. Pigou, 1877-1959). If the polluter has to pay a per-unit tax of $t = e'(x^o)$, then the invisible hand would again restore efficiency, as illustrated in Figure 1.

The points made here for a negative production externality apply, with suitable modifications, to other non-pecuniary externalities. Take one case, e.g., positive production

externality, and see how far you can get. E.g., the Pigouvian tax becomes a subsidy, to boost production to the socially optimal level.

Pecuniary externalities are often regarded as wealth transfers, not as sources of inefficiency. This is debatable, but I will not do so here.

VI. **Network externalities.** Sometimes the value of a good depends on how many others use it. For example,

- Facebook is valuable to consumers mainly because so many other consumers use the platform. Likewise for popular on-line games.
- The QWERTY keyboard is valued mainly because so many people already use it. Conversely, US firms refuse to switch to the metric system because it isn't used by very many of their suppliers and customers.
- High trading volume lowers transactions costs on the New York Stock Exchange, as compared to a startup trading platform.
- Classic examples include narrow-gage railroads, telephone service and fax machines. The internet is a very prominent example in the 21st century.
- A classic "folk" model is that the costs of a network are proportional to the number of agents n, but benefits are proportional to the number of possible pairs of agents, roughly n^2.
- Thus we have increasing returns to scale, and a natural monopoly. The network effects are a sort of barrier to entry.

Suppose a new platform (or game or app) comes along that, given the same level of usage, would be more efficient than the incumbent.

- If strong network effects are present, the new platform may never get enough usage to catch on, which would be a loss of TS.
- Some observers argue that that sort of inefficient "lock-in" is common, and exemplified in the failed betamax format for video casette recorders and the Dvorak

keyboard. If so, the invisible hand has gone astray. (Others dispute that, and argue that the VHS and QWERTY formats are more efficient than claimed.)

- Lab experiments with weakest link games are indisputable instances.
- Managerial economics texts argue for "penetration pricing," bring-a-friend bonuses, and other up-front inducements to overcome the problem.

VII. **VCG and other demand-revealing mechanisms.** Recall that Lindahl pricing, Pigouvian taxes and Coasian bargaining all assume that agents know each others' WTP. Of course, true WTP is typically unknown by other parties. How can this informational problem be overcome?

The problem arises in many contexts, not just for externalities. We first saw it when considering price discrimination, and noted that self-sorting mechanisms (as in 2^opd) can mitigate the problem.

Another place it shows up is in auction theory. The Vickrey second price auction encourages bidders to fully reveal WTP.

Varian presents Clarke-Groves mechanisms for public good provision.

The basic ideas in these Vickrey-Clarke-Groves (VCG) mechanisms are

- Arrange things so that announcing true WTP is a dominant strategy.
- This requires that the payment each agent receives (or pays) is independent of her announced WTP, but the outcome (e.g., winning an auctioned item, or the amount Y of public good) does depend on her announcement.
- The payment is often calculated as the agent's externality, i.e., $p_i = [TS|-i]_{-i} - [TS|i]_{-i}$, where
- $[TS|-i]_{-i}$ (resp $[TS|i]_{-i}$) is total surplus of agents other than i when i does not (resp does) participate.

Example: Ad auctions, GSP vs VCG.