Lecture Notes for Economics 200C: Games and Information
Vincent Crawford, revised March 2000; do not reproduce except for personal use

1. Introduction

MWG 217-233; Kreps 355-384; Varian 259-265; McMillan 3-41
Robert Gibbons, "An Introduction to Applicable Game Theory," *Journal of Economic Perspectives* (Winter 1997), 127-149 (or you can substitute his book)

Noncooperative game theory tries to explain outcomes (including cooperation) from the basic data of the situation, in contrast to *cooperative* game theory, which assumes unlimited communication and cooperation and tries to characterize the limits of the set of possible cooperative agreements. In "parlor" games players often have opposed preferences; such games are called *zero-sum*. But noncooperative game theory spans the entire range of *interactive decision problems* from pure conflict to pure cooperation (coordination games); most applications have elements of both.

A *game* is defined by specifying its *structure*: the players, the "rules" (the order of players' decisions, their feasible decisions at each point, and the information they have when making them); how their decisions jointly determine the outcome of the game; and their preferences over possible outcomes. Any uncertainty about the outcome is handled by assigning *payoffs* (von Neumann-Morgenstern utilities) to the possible outcomes and assuming that players' preferences over uncertain outcomes can be represented by expected-payoff maximization.

Assume game is a complete model of the situation; if not, make it one, e.g. by including decision to participate. Assume numbers of players, decisions, and periods are finite, but can relax as needed.

Something is *mutual knowledge* if all players know it, and *common knowledge* if all know it, all know that all know it, and so on. Assume common knowledge of structure (allows uncertainty with commonly known distributions modeled as "moves by nature"): games of *complete information*.

Can represent a game by its *extensive form* or *game tree.* E.g. contracting by ultimatum (MWG uses Matching Pennies, Kreps has abstract examples): Two players, R(ow) and C(olumn); two feasible contracts, X and Y. R proposes X or Y to C, who must either accept (a) or reject (r). If C accepts, the proposed contract is enforced; if C rejects, the outcome is a third alternative, Z. R prefers Y to X to Z, and C prefers X to Y to Z. R's preferences are represented by *vN-M utility* or *payoff* function u(y)=2, u(x)=1, u(z)=0; and C's preferences by $v_2(x)=2$, $v_2(y)=1$, $v_2(z)=0$.

Draw game trees when C can observe R's proposal before deciding whether to accept, and when C cannot. Order of *decision nodes* has some flexibility, but must respect timing of information flows. Players assumed to have *perfect recall* of their own past moves and other information; tree must reflect this. Each decision node belongs to an *information set*, the nodes the player whose decision it is cannot distinguish (and at which he must therefore make the same decision). All nodes in an information set must belong to the same player and have the set of same feasible decisions. Identify each information set by circling its nodes (MWG) or connecting them with dotted lines (Kreps).

A *static game* has a single stage, at which players make simultaneous decisions, as in contracting with unobservable proposal. A *dynamic game* has some sequential decisions, as in contracting with observable proposal. A game of *perfect information* is one in which a player making a decision can always observe all previous decisions, so every information set contains a single decision node, as in contracting with observable proposal. Complete does not imply perfect information; but if background uncertainty is modeled as "moves by nature," perfect implies complete information.

A *strategy* is a complete contingent plan for playing the game, which specifies a feasible decision for each of a player's decision nodes in the game and possible information states when he reaches them. (In static games I sometimes say "decision" or "action" instead of "strategy.") A strategy is like a detailed chess textbook, *not* like a move. A player's feasible strategies must be independent of others' strategies (no "wrestle with the other cricket"), and specifying a strategy for each player determines an outcome (or at least a probability distribution over possible outcomes) in the game.

In contracting, whether or not C can observe R's proposal, R has two pure strategies, "(propose) X" and "(propose) Y." If C cannot observe R's proposal, C has two pure strategies, "a(ccept)" and "r(eject).". If C can observe R's proposal he can make his decision depend on it, and therefore has four pure strategies, "a (if X proposed), a (if Y proposed)", "a, r", "r, a", and "r, r."

The above descriptions apply to *mixed* strategies (randomized choices of pure strategies) as well as *pure* (unrandomized) strategies. In games with perfect recall mixed strategies are equivalent to *behavior* strategies, probability distributions over pure decisions at each node (Kuhn's Theorem).

A strategy must be a complete contingent plan (even for nodes ruled out by own prior decisions!) so that in dynamic games we can evaluate the consequences of alternative strategies, to formalize the idea that the predicted strategy choice is optimal. This is a surprising difference from individual decision theory, where zero-probability events can be ignored. In games we must pay attention to zero-probability outcomes because they are *endogenously* determined by players' decisions.

Because strategies are complete contingent plans, players must be thought of as choosing them *simultaneously* (without observing others') at the start: theory assumes rational foresight, so simultaneous choice of strategies is same as decisions in "real time."

A *game* maps strategy profiles into payoffs; a *game form* maps profiles into outcomes, without specifying payoffs. The relationship between strategy profiles, outcomes, and payoffs is often described by the *normal form* or the *payoff matrix* or *payoff function*.

In contracting, without and with observable proposals, the payoff matrices are:

	a	r
X	2 1	0 0
Y	1 2	0 0

Contracting with
Unobservable Proposal

	a, a	a, r	r, a	r, r
X	2 1	2 1	0 0	0 0
Y	1 2	0 0	1 2	0 0

Contracting with
Observable Proposal

Increasingly game-theoretic examples illustrate some issues that a theory of games should address (see also abstract, mostly normal-form examples at Kreps 389-392).

	L	R
T	2 2	1 2
B	2 1	1 1

Crusoe "versus" Crusoe

	Confess	Don't
Confess	-5 -5	-10 -1
Don't	-1 -10	-2 -2

Prisoner's Dilemma

	Push	Wait
Push	1 5	5 3
Wait	-1 9	0 0

Pigs in a Box

	Heads	Tails
Heads	-1 1	1 -1
Tails	1 -1	-1 1

Matching Pennies

Crusoe versus Crusoe is not really a game, just two individual decision problems; each player necessarily has best (dominant) strategy, independent of the other's strategy.

In Prisoner's Dilemma, players' decisions affect each other's payoffs, but each player has dominant strategy. But because of payoff interactions, these individually optimal decisions yield collectively suboptimal (inefficient, in terms of their preferences alone) outcome.

In Pigs in a (Skinner) Box, Row (R) is a dominant (big) pig and Column (C) a subordinate (little) pig. There is a lever at one end, which when pushed yields 10 units of grain at the other end. Pushing costs a pig the equivalent of 2 units of grain. If the big pig pushes while the little pig waits, the little pig can eat 5 units before the big pig gets there and shoves him aside. If the little pig pushes while the big pig waits, the little pig cannot push the big pig aside and the big pig gets all but one unit. If both push, arriving at the grain at the same time, the little pig gets 3 units and the big pig gets 7 units. If both wait, both get 0.

When behavior settles down in experiments with real pigs, it tends to be at R Push, C Wait. The little pig (C) does better, even though R can do anything C can do and must therefore do at least as well in any individual decision problem (or in any zero-sum game)! Examining the game suggests that this happens because for C, but not for R, Wait dominates Push, so that only R has an incentive to Push. Evidently, in games weakness can be an advantage! R would do better if he could commit himself, say by limiting his ability to shove C aside, to giving C some of the grain if C Pushed. We would like to know which kinds of games such commitments help in, and what kinds of commitments help. Understanding this would help in understanding the usefulness of contracts.

Matching Pennies has no good pure strategies, but a unique good way to play using mixed strategies: shows strategic importance of mixed strategies in conflict situations.

The abstract 3x3 game has a unique profile of pure strategies such that each player's strategy is best for him, given the other's, but no dominance. (It is easy to show from best-response cycles that it has no mixed-strategy equilibria.) It shows the need for a way of analyzing players' strategy choices that takes their interdependence fully into account.

Alphonse and Gaston's problem is that there are *two* ways to solve their coordination problem. Each is best for both if both expect the game to played that way, but not otherwise; and each requires them to behave differently when there are no clues they can use to distinguish their roles (as they must, to behave in systematically different ways). This illustrates the nontrivial problems players may face even when their preferences over outcomes are the same, or nearly the same. We say economics is about coordination, but the usual analysis *assumes* coordination by assuming competitive equilibrium (Walrasian auctioneer). This is okay for some purposes, but leaves us no way to think about the influence of the environment on coordination; e.g., it's just as easy to imagine a million people's decisions perfectly coordinated as two people's, so we have little or no theory of the effect of group size on the efficiency of coordination. Battle of the Sexes complicates Alphonse and Gaston's problem with different preferences about how to coordinate, but still no clues about how to break the symmetry as needed for efficient coordination. (The problem is clearer if Battle of the Sexes is transposed to Hawk-Dove game, which uses decision labels that identify strategies that have same meaning in terms of the structure.)

	L	C	R
T	0 7	5 0	7 0
M	0 5	2 2	0 5
B	7 0	5 0	0 7

Unique Profile of Mutually Best but Not Obviously Best Pure Strategies

	Go	Wait
Go	0 0	1 1
Wait	1 1	0 0

Alphonse and Gaston

	Fights	Ballet
Fights	1 2	0 0
Ballet	0 0	2 1

Battle of the Sexes

	Other	Player
	Stag	Rabbit
Stag	2 2	1 0
Rabbit	0 1	1 1

Two-Person Stag Hunt

	All Other	Players
	All-Stag	Not All-Stag
Stag	2	0
Rabbit	1	1

n-Person Stag Hunt

	L	C	R
T	10 4	0 3	3 1
B	0 0	10 2	3 10

Domination Via Mixed Strategies

	L	C	R
T	0 7	5 0	3 0
M	0 5	2 2	0 5
B	7 0	5 0	3 7

Dominance-Solvable

	L	R
T	1 1	1 0
B	0 0	0 0

Give Me a Break

	L	R
T	1 1	0 0
B	0 0	0 0

Give Us a Break

In Stag Hunt (Rousseau's story, faculty meeting, assembly line) all-Stag is better for all players than all-Rabbit; but unless all others play Stag, a player does better with Rabbit. With two or n players, there are two symmetric, Pareto-ranked, pure-strategy equilibria (and an uninteresting mixed-strategy equilibrium). All-Stag is the "obvious" solution, but playing Stag is risky because it may not be sufficiently clear that all-Stag is obvious to all, or sufficiently clear that it is sufficiently clear, etc. Rabbit is safer because its payoff, although lower than Stag's when all play Stag, is independent of others' decisions. Stag seems a good bet if there are only a few people, but not if there are ten, or a hundred. This game poses a different kind of coordination problem, with a tension between the higher payoffs of all-Stag and its greater fragility. Stag versus Rabbit is like the choice between participating in a complex and highly productive but fragile society and autarky, which is less rewarding but also less dependent on coordination.

Problems 7.C.1, 7.D.1-2, and 7.E.1 at MWG 233-234 (all answered in manual); and problems 3-4 at Kreps 385-386.

2. Games with Simultaneous Moves ("Static Games")

MWG 235-253, 258-261, and 387-400; Kreps 387-417, 437-443, and 551-556; Varian 265-268

Assume common knowledge of structure and players' rationality, in the standard decision-theoretic sense of choosing strategies that maximize expected payoff given some *beliefs* about others' strategies not contradicted by anything he knows and following Bayes' Rule.

Define *strictly dominant* and *dominated* strategies. Dominance for pure implies dominance for mixed strategies, but can have dominance by mixed without dominance by pure strategies. Define *weak dominance*. Prisoner's Dilemma, Pigs in a Box, Domination via Mixed Strategies. Define *iterated deletion* of strictly dominated strategies ("iterated dominance"). Independence of order of elimination (proof MWG 262, problem 8.B.4), but independence doesn't generalize to iterated weak dominance (e.g. Give Me a Break). Define *dominance-solvability*, e.g. in 3x3 example above.

Rationalizable strategies survive the iterated removal of strategies that are never a *(weak) best response* (MWG 242-243). Close to iterated *strict* dominance. Order of removal of never-weak-best-response strategies doesn't matter (MWG 262, problem 8.C.2).

In contracting example, C should accept whether or not C can observe R's proposal, and if so, whichever contract R proposes. R should therefore propose Y, his most preferred contract. When R's proposal is unobservable, C's strategy a dominates r; given that, R's strategy Y dominates X. This leaves Y and a as the only rationalizable strategies.

The set of rationalizable strategies can't be larger than the set that survive iterated strict dominance, because strictly dominated strategies can never be weak best responses (can't extend to iterated weak dominance, e.g. Give Me a Break). In two-person games the two sets are the same because never-weak-best-response strategies are exactly those that are strictly dominated. *Any* strategies in Matching Pennies, Battle of the Sexes, or 3x3 game with unique pure equilibrium, no dominance is rationalizable. With more than two persons, the set of rationalizable strategies can be smaller than the set that survives iterated strict dominance (MWG Exercise 8.C.4 at 245; Brandenburger 87-88).

Rationalizability characterizes implications of common knowledge of structure, rationality:

Theorem: Common knowledge of the structure and rationality implies players will choose rationalizable strategies, and any profile of rationalizable strategies is consistent with common knowledge of the structure and rationality. Proof (MWG 243): Illustrate in first direction for 3x3 dominance-solvable game, second direction by building tower of beliefs to support corners in 3x3 game with unique pure-strategy equilibrium but no dominance.

There is a link between number of rounds of iterated dominance and of iterated knowledge of rationality; need *common* knowledge only for indefinitely large games.

Most important games (and all coordination games) have multiple rationalizable outcomes, so players must base their decisions on predictions of others' decisions not dictated by common knowledge of rationality. This prediction is interdependent: "theory of interdependent decisions." In games with multiple rationalizable outcomes, much of the theory's power comes from assuming players choose strategies in *Nash equilibrium*, a strategy profile for which each player's strategy is a *best response* to other players' strategies (a fixed point of the best-response correspondence).

A Nash equilibrium is a kind of "rational expectations" equilibrium, in that if all players expect the same strategy profile and choose strategies that are best responses given their beliefs, beliefs will be confirmed iff they are in equilibrium. (Differs from usual rational expectations in that individuals' decisions are predicted, and players' predictions interact.)

Nash equilibrium is often *identified* with "rationality," but equilibrium is much stronger than common knowledge of rationality. Any equilibrium strategy is rationalizable, but equilibrium also requires players' strategies to be best responses to *correct* beliefs about others' strategies (which must then be the same for all), not just some beliefs consistent with common knowledge of rationality. E.g., belief towers supporting non-equilibrium strategy profiles in 3x3 game with unique pure-strategy equilibrium but no dominance.

Unlike rationalizability, equilibrium is a property of strategy *profiles*, relationship between strategies. "Equilibrium strategy" often refers to any strategy that's part of *an* equilibrium.

In contracting example when R's proposal is observable, only r, r for C is strictly dominated. When r, r is eliminated there is no strict dominance, so remaining strategies are all rationalizable. Common-sense outcome, (Y; a, a), is an equilibrium, but there are two others, (Y; r, a) and (X; a, r), one with outcome X! In these equilibria C plans to reject one of R's possible proposals, irrationally, and R's anticipation of this keeps him from making that proposal, so the irrationality does not reduce the payoff of C's strategy in the entire game. Such equilibria are said to involve "incredible threats" (misleadingly because not explicit like a real threat, but implicit in expectations that support equilibrium).

Theorem: Equivalence of iterated strict dominance and equilibrium in dominance-solvable games. Not of iterated weak dominance, order of elimination matters, Give Me a Break.

Consistency of iterated strict dominance and equilibrium in non-dominance-solvable games. Nothing that doesn't survive iterated strict dominance can be in an equilibrium.

An equilibrium can be either in pure or mixed strategies. A mixed strategy profile is an equilibrium if for each player, his mixed strategy maximizes his expected payoff over all feasible mixed strategies (MWG Dn. 8.D.2 at 250), e.g. Matching Pennies, where uncertainty is purely strategic. (Nash's population interpretation, "beliefs" interpretation.)

Theorem: A mixed strategy profile is an equilibrium if and only if for each player, all pure strategies with positive probability yield the same expected payoff, and all pure strategies he uses with zero probability yield no higher expected payoff (MWG Prop. 8.D.1 at 250-251). (Kuhn-Tucker conditions for maximizing expected payoffs linear in probabilities.)

Theorem: Every finite game has a mixed-strategy Nash equilibrium (Prop. 8.D.2 at MWG 252).

Existence of equilibrium may require mixed strategies, in general.

The next theorem gives a more abstract and more general existence result.

Theorem: Every game whose strategy spaces are nonempty, convex, and compact subsets of Euclidean space, and whose payoff functions are jointly continuous in all players' strategies and quasiconcave in own strategies has a Nash equilibrium (Prop. 8.D.3 at MWG 253).

Interpret for mixed strategies in finite games, pure strategies in games with continuously variable pure strategies. Theorem also implies the existence of rationalizable strategies.

Nonuniqueness of equilibrium and refinements. E.g. Give Us a Break, Contracting with observable proposal, discrete Nash demand game (Kreps 551-556).

A Nash equilibrium is (normal-form) *trembling-hand perfect* if there is some sequence of tremble-perturbed games converging to the game for which there is some sequence of Nash equilibria converging to the equilibrium (Dn. 8.F.1 at MWG 258). E.g. Give Me a Break, Give Us a Break.

Theorem: A Nash equilibrium is trembling-hand perfect iff there is a sequence of totally mixed strategies converging to the equilibrium such that each player's equilibrium strategy is a best response to every element of the sequence (Prop. 8.F.1 at MWG 259).

Theorem: In a trembling-hand perfect equilibrium, no weakly dominated strategy can be played with positive probability (Prop. 8.F.2 at MWG 259). Any strict equilibrium (define) is trembling-hand perfect. Any finite game has a trembling-hand perfect equilibrium.

Selection among strict equilibria: Harsanyi-Selten's General Theory, risk- and payoff-dominance in two-person and n-person Stag Hunts.

Rationale: Equilibrium is at least a necessary condition for a prediction about behavior in a game if players are rational and there is to be a unique prediction. But how might players come to have correct beliefs about how a game will be played? The answer given to this question separates two leading approaches to noncooperative game theory:

(i) the traditional, deductive approach assumes that players independently deduce correct beliefs from common knowledge (why *common*?) of a theory of strategic behavior that makes a unique prediction for the game in question, like Harsanyi and Selten's General Theory of Equilibrium Selection; such expectations must yield an equilibrium, given rationality (proof);

(ii) adaptive learning approach, based on common understanding of previous analogous games, in which players' strategies adjust over time in response to observed payoffs.

These approaches agree that possible limiting outcomes are something like Nash equilibria (in the game that is repeated, not the game that describes the entire process); but they generally differ on convergence and the likelihood of alternative equilibria.

Problems 8.B.1-7, 8.C.1-4, and 8.D.1-9 at MWG 262-266, problems 12.C.1-17 at MWG 430 (answered in manual); and problems 2-3, 14-15, 17-18, and 21 at Kreps 451-462.

3. Games with Sequential Moves ("Dynamic Games")

MWG 267-282, 405-417, and 423-427; Kreps 401-402, 417-449, and 556-565; Varian 273-278

Display Examples 9.C.1 at MWG 283, 9.B.3 at 274, 9.C.4 at 289, 9.C.5 at 290, 9.B.5 at 282, and Figure 9.D.1 at 293.

In dynamic games some useful ideas depend on extensive form. E.g. in contracting with observable proposal, the common-sense outcome (Y; a, a) is an equilibrium, but there are two other equilibria, (Y; r, a) and (X; a, r) with "incredible threats," but which survive iterated strict dominance (cf. "predation game" Example 9.B.1 at MWG 268-269). More generally, whenever play doesn't reach a given node in an equilibrium, equilibrium doesn't restrict the decision at that node at all.

Sequential rationality and time-consistency of "solution concept" (on and off equilibrium path, unlike decision-theoretic notion). Relation to iterated *weak* dominance, equilibrium in un-weakly dominated strategies, in contracting example (cf. Example 9.B.2 at MWG 271).

Zermelo's Theorem (Proposition 9.B.1 at MWG 272-273): Backward induction, existence, and (if no ties) uniqueness of sequentially rational strategies in finite games of perfect information.

Extension of sequential rationality/backward induction to finite games of *imperfect* information via plugging in payoffs of equilibria in subgames and folding back, e.g. predation game with simultaneous choices following entry but unique equilibrium (Example 9.B.3 at MWG 273-274).

A *subgame* is a subset of a game that starts with an information set with a single node, contains all and only that node's successors in the tree, and contains all or none of the nodes in each information set (Definition 9.B.1 at MWG 274). A *subgame-perfect equilibrium* is a strategy profile that induces an equilibrium in every subgame (MWG 275, Dn. 9.B.2).

Theorem (Proposition 9.B.2 at MWG 276, generalizes Zermelo's Theorem, proof the same): Existence of pure-strategy subgame-perfect equilibrium, and (if no ties) uniqueness of subgame-perfect equilibrium in finite games of perfect information. Existence of subgame-perfect equilibrium in games of imperfect or incomplete information (Example 9.B.3 at MWG 273-274).

Trembling-hand perfect equilibria are subgame-perfect, but not vice versa.

Theorem (Proposition 9.B.3 at MWG 277-278): Characterization of subgame-perfect equilibria via backward induction.

Theorem (Proposition 9.B.4 at MWG 279-280): Equivalence of subgame-perfect equilibrium to concatenated equilibria of period games in finite-horizon games (even with imperfect information) with unique equilibria and immediate observability of pure strategies each period, payoffs summed over periods. Illustrate proof in finitely repeated Prisoner's Dilemma.

Theorem (Proposition 9.B.9 at MWG 302): With multiple period equilibria or infinite horizon can get nonconcatenated subgame-perfect equilibria.

Subgame perfect equilibrium as inadequate formalization of sequential rationality, Example 9.C.1 at MWG 282-283 (one choice by first mover leads to two-node information set, no subgames).

A *system of beliefs* is a probability distribution over nodes, which gives the relative likelihoods of being at each node in an information set, conditional on having reached it (Dn. 9.C.1 at MWG 283).

A strategy profile is *sequentially rational* at an information set if no player can do better, given his beliefs about what has happened so far, by changing his strategy (Dn. 9.C.2 at MWG 284). Generalizes above notion of sequential rationality to games like Example 9.C.1 at MWG 282-283.

A strategy profile and system of beliefs is a *weak perfect Bayesian equilibrium* if the strategy profile is sequentially rational given the beliefs, and the beliefs are derived from the strategy profile using Bayes' Rule whenever possible (Dn. 9.C.3 at MWG 285, Example 9.C.1 at MWG 282-283). ("Weak" because completely agnostic about zero-probability updating.)

Theorem (Proposition 9.C.1 at MWG 285-286): A strategy profile is an equilibrium in an extensive form game if and only if there exists a system of beliefs such that the profile is sequentially rational given the beliefs at all information sets that have positive probability of being reached by the profile, and beliefs are derived from profile using Bayes' Rule whenever possible.

A strategy profile and system of beliefs is a *sequential equilibrium* if the profile is sequentially rational given the beliefs, and there exists a sequence of completely mixed strategies converging to the profile, such that the beliefs are the limit of beliefs derived using Bayes' Rule from the totally mixed strategies (Dn. 9.C.4 at MWG 290). Strengthens weak perfect Bayesian by requiring more consistency of zero-probability beliefs, adding equilibrium play off equilibrium path. Closely related to *perfect Bayesian equilibrium*, defined at MWG 452.

A sequential equilibrium is (trivially) a weak perfect Bayesian equilibrium, but not vice versa.

Theorem (Proposition 9.C.2 at MWG 291, Example 9.C.1 at MWG 282-283): A sequential equilibrium is subgame-perfect, but not vice versa.

Centipede Game and critique of sequential rationality (MWG 281-282, Kreps 401-402).

More refinements:

Forward induction, relation to iterated *weak* dominance, e.g. in Battle of the Sexes with outside option; ambiguity of notion in general (MWG 292-296, Figure 9.D.1 at 293). Powerful, tricky.

Extensive-form trembling-hand perfect equilibrium as trembling-hand perfect equilibrium in *agent normal form* (Dn. 9.BB.1 at MWG 299-301).

Analysis of time-sequenced strategy choices with and without observability, irreversibility: without observability, order of moves doesn't alter feasible strategies, payoffs, normal form, or (subgame-perfect) equilibrium outcomes. Without irreversibility (or costly reversibility, in which case decision to incur costs is irreversible) a similar conclusion holds. In particular, announcement with no direct payoff effect ("cheap talk") of "intention" to choose a strategy has no effect, although it could focus beliefs on a particular subgame-perfect equilibrium in Alphonse-Gaston or Stag Hunt.
Importance of separating assumptions about structure and behavior. Different "solution concepts" as same behavioral assumptions in different games, e.g. Stackelberg, Cournot, and Bertrand.

Theorem (Proposition 12.C.1 at MWG 388): Bertrand duopoly with constant returns to scale, perfectly substitutable goods: Simultaneous price choices by firms yields competitive outcome as unique equilibrium. No calculus despite continuously variable strategies because of discontinuities.

Theorem (Proposition 12.C.2 and Example 12.C.1 at MWG 390-392): Cournot duopoly with constant returns to scale, perfectly substitutable goods: Simultaneous quantity choices by firms yields equilibrium (not necessarily unique) with prices between competitive and monopoly prices.
Equilibrium via calculus at last!

Cournot outcome approaches competitive outcome as number of firms grows (MWG 391-394, Kreps 443-449).

Stackelberg leadership as subgame-perfect equilibrium with sequenced quantity choices by firms.

Capacity constraints and product differentiation (refer to MWG 394-400). Kreps-Scheinkman and importance of timing of irreversible decisions.

Simultaneous entry decisions followed by Bertrand or Cournot competition (refer to MWG 405-411).

Strategic precommitments, strategic complements and substitutes, direct and indirect effects of investment decisions (MWG 414-417).

Three views of entry deterrence: irreversible decisions that affect future interactions, repeated games (Section 4), informational (Section 5).

Dixit and Spence entry deterrence and accommodation models (MWG 423-427).

Problems 9.B.1-11 and 14 at MWG 301-305, problem 12.BB.1 at MWG 427, problem 12.C.18 at MWG 433, problems 12.E.1-7 and 12.F.1-4 at MWG 434-435; and problems 1, 4, 16, and 20-21 at Kreps 451-462

4. Infinite Horizon and Repeated Games

MWG 296-299, 400-405, and 417-423; Kreps (better on this topic) 503-515, 524-526, and 551-565; Varian 269-271

A. Complete information alternating-offers bargaining models

Two players bargain via alternating offers over v dollars. (Relation between discount factors and discount rates.) Player 1 (chosen arbitrarily) begins in period 1 with a continuously variable offer between 0 and v. If player 2 accepts he gets the offered amount and player 1 gets what's left of v (assumes none of the money is wasted, but can relax). If player 2 rejects and there are any periods remaining, player 2 makes a counteroffer in period 2, and so on. The value of any agreement is discounted by a common (can relax) discount factor $0 \leq \delta \leq 1$, so delay is costly.

Ultimatum Game (one-period version of alternating-offers bargaining): unique subgame-perfect equilibrium in which player 1 ("proposer") makes a proposal in which he gets all of the surplus and player 2 ("responder") accepts (despite indifference). Leader gets all of the surplus, therefore has incentive to structure his proposal to maximize it so the outcome is efficient (Edgeworth Box).

Reinterpretation of this equilibrium as limit of (risk-dominant) subgame-perfect equilibria without indifference in discrete-offer version (still have equilibrium with indifference too). Ultimatum Game often generalized to allow nontrivial structuring of proposals and used as model of bargaining over contracts. Graphical analysis of subgame-perfect equilibrium in utility-possibility set with downward-sloping frontier (link to underlying game).

Probably okay for most purposes, but use with care because predictions don't do well when Ultimatum Game is played in laboratory (Responders get upset enough to reject low but positive offers; Proposers anticipate this, and respond by offering much more than nothing).

Pareto-efficiency of subgame-perfect equilibrium outcome. "Noncooperative" contracting game yields "cooperative" equilibrium outcome, in this case because in equilibrium the Proposer gets all of the surplus from an agreement, and therefore has an incentive to propose an agreement that maximizes it. Model explains Pareto-efficiency of outcome rather than just assuming it. Typical of literature on incentives and mechanism design.

Finite-horizon alternating-offers bargaining (I. Stahl): unique subgame-perfect equilibrium in which player 1 makes a proposal in which he gets all of the surplus from reaching an agreement immediately, relative to an agreement delayed by one period, and player 2 accepts (again despite indifference). Two-period example with $v = 1$, $\delta = ¾$.

Subgame-perfect equilibrium outcome is again Pareto-efficient, because player 1 again has an incentive to make a proposal that maximizes the surplus (though trivial structuring of proposals in the simple model trivializes this issue) and there's no delay in equilibrium.

Surplus-sharing is entirely determined by delay costs, with first-mover advantage for player 1 when there is an odd number of periods (even though the choice of first-mover was arbitrary). First-mover advantage goes away as δ approaches 1 and the horizon approaches infinity. In the limit, with equal discount factors, surplus-sharing approaches equal split (with a general, convex utility-possibility set, approaches the Nash bargaining solution, which generalizes equal-split to such sets).

Predictions again don't do well when alternating-offers bargaining games are played in laboratory, partly because Responders get upset as in Ultimatum Game, and partly because the longer the horizon the more complex the backward induction/iterated dominance argument required to identify the subgame-perfect equilibrium, and real people don't believe that others will follow it.

Infinite horizon version (reinterpretation as *potentially* infinite horizon, with conditional probabilities of continuation bounded above zero and perhaps discounting too; Rubinstein a982 *EMT*; MWG 296-299, Kreps 556-565): unique subgame-perfect equilibrium in which player 1 makes a proposal in which he gets all of the surplus from reaching an agreement immediately, relative to an agreement delayed by one period (anticipating subgame-perfect equilibrium in subgame following rejection), and player 2 accepts (again despite indifference). Sketch proof.

Subgame-perfect equilibrium outcome is Pareto-efficient, for same reasons, equals limit of finite-horizon subgame-perfect equilibria as horizon approaches infinity.

Surplus-sharing still entirely determined by delay costs, with first-mover advantage for player 1. As δ approaches 1 the first-mover advantage goes away, and with equal discount factors approaches equal split (or Nash (1950 *EMT*) bargaining solution in more general games).

Game has a continuum of Nash equilibria without subgame-perfectness. Perfectness has force here because of the way it interacts with the alternating-offers rules and delay costs.

By far most popular bargaining model, but result doesn't generalize to n players, discrete offers, incomplete information, almost-common knowledge of rationality (Kreps 552-565; Kreps, *Game Theory and Economic Modelling*); and doesn't do well in experiments.

Alternatives:

With no delay cost, fixed horizon, and fixed pattern of alternating offers, the *last* mover gets all of the surplus, in effect becoming the Proposer in an Ultimatum Game by unilaterally forcing delay (Kreps 551-556).

When strategic uncertainty and risk of coordination failure are more important than delay costs, there's a fixed horizon, but there is no fixed pattern of alternating offers ("lock 'em in a room" bargaining), the analysis is very different. Nash's (1953 *EMT*) demand game with strategies viewed as the least surplus each player can be induced to accept. Role of expectations, culture, focal points, strategic moves in determining bargaining outcomes. Nash's axiomatic (1950 *EMT*) solution.

B. Complete-information repeated games

Define *repeated game* as dynamic game in which same *stage game* is played over and over again by same players. Stage game could be anything, even another repeated game.

Repeated Prisoner's Dilemma (location of apostrophe, methodological individualism). Canonical (but overworked, and not entirely representative) model of using repeated interaction to overcome short-run incentive problems that work against efficiency.

With finite horizon (however large) "Defect-Defect no matter what" is unique subgame-perfect equilibrium. No cooperation on equilibrium path in any equilibrium (Kreps 514).

With infinite horizon and a sufficiently high discount factor, trigger strategies ("Cooperate till other guy Defects, then Defect forever") support cooperation on equilibrium path. Okay, but not perfect. Modified trigger strategies ("Cooperate till either of us Defects, then Defect forever") support cooperation as subgame-perfect equilibrium. "Implicit contract." No true altruism, just "reciprocal altruism" supported as subgame-perfect equilibrium with purely self-interested behavior.

Subgame-perfectness is important here because without it, players wouldn't want to carry out the punishments, and their anticipations of this would render planned punishments ineffective. Even with subgame-perfectness there's a potential problem with renegotiation because punishments are inefficient. Renegotiation-proofness kills possibility of supporting cooperation in repeated Prisoner's Dilemma, limits it elsewhere.

	Cooperate	**Defect**
Cooperate	3 3	5 0
Defect	0 5	1 1

With these payoffs, players can support repeated Cooperate-Cooperate in subgame-perfect equilibrium using above modified trigger strategies (or any strategies) if and only if $3(1 + \delta + \delta^2 +\ldots) = 3/(1 - \delta) \geq 5 + 1(\delta + \delta^2 +\ldots) = 5 + \delta/(1 - \delta)$ if and only if $\delta \geq ½$.

When $\delta \leq ½$, the future is not important enough for threats of future defection to support cooperation, and only repeated Defect-Defect is consistent with subgame-perfect equilibrium (or equilibrium). The limit of ½ is dependent on the magnitudes of the payoffs in the example, but fact that higher values of δ never hurt is general.

In the infinite-horizon repeated Prisoner's Dilemma, there are also many asymmetric subgame-perfect equilibria. Example (Kreps 507): Implicit contract is Row alternates between Cooperate and Defect and Column always Cooperates. This continues until either deviates, in which case both Defect from then on. Column does worse, and threat is symmetric, so supporting Column's strategy as part of a subgame-perfect equilibrium is unambiguously harder than supporting Row's strategy. In the hypothesized equilibrium Column gets $3 + 0\delta + 3\delta^2 +\ldots = 3/(1 – \delta^2) \geq 5 + 1(\delta + \delta^2 +\ldots) = 5 + \delta/(1 - \delta)$ if and only if $\delta \geq 0.59$ (approximately), so the asymmetric implicit contract is consistent with subgame-perfect equilibrium as long as $\delta \geq 0.59$. The limit is higher than for the symmetric implicit contract because the asymmetry makes it harder to keep both players willing to stay with the contract when the alternative is a symmetric punishment.

A general feature of infinite-horizon repeated games is an enormous multiplicity of equilibria, multiplicity both of equilibrium outcomes and the threats that can be used to support them (which in this noiseless version of the game never need to be carried out on the equilibrium path).

The perfect symmetry of the repeated Prisoner's Dilemma makes it seem easy to choose an equilibrium to represent its implications. However, even here the symmetric Pareto-efficient outcome is supported by threats to totally destroy the relationship if anything goes wrong. This need for near-perfect coordination of strategy choices makes the implicit contract very fragile. (And often players only get to play the game once, so learning justifications for equilibrium may not be available.) More "forgiving" strategies are less effective in deterring cheating. In real environments there is a tension between efficiency and fragility, which is not yet at all well understood.

We've seen one symmetric and one asymmetric efficient equilibrium of the repeated Prisoner's Dilemma. It would be useful to know, more generally, what kinds of implicit contracts can be supported as subgame-perfect equilibria in repeated games. Define *minimax* (not maximin) payoff: minimum over others' strategies of maximum of own payoff over own strategy given others' strategies. E.g. Prisoner's Dilemma, Bertrand, Cournot, and Stackelberg duopoly.

Folk Theorem: In an infinitely repeated game with complete information and observable strategies, for any feasible pair of payoffs strictly greater than those that follow from repeating players' minimax payoffs in the stage game, there is a discount factor such that for all greater discount factors, those payoffs arise in a subgame-perfect equilibrium of the repeated game (Prop. 12.AA.5 at MWG 422, stated in terms of average payoffs, Example 12.AA.1; Kreps 508-509).

Easy to prove as in Prisoner's Dilemma for Nash reversion. Harder for minimax. Temporal convexification with high discount factors.

Application to oligopoly (real meaning of reaction function is kind of strategy in repeated oligopoly, conceptually (not mathematically) distinct from best-response function).

Theorem (Proposition 12.D.1 at MWG 401; implicit collusion in infinitely repeated Bertrand duopoly): With a sufficiently high discount factor, the monopoly price can be supported as a subgame-perfect equilibrium outcome in an infinitely repeated Bertrand duopoly by threats to revert forever to the competitive price if anyone deviates.

Theorem (Proposition 12.D.2 at MWG 403; Folk Theorem in infinitely repeated Bertrand duopoly): With a sufficiently high discount factor, any price from the competitive to the monopoly price can be supported as a subgame-perfect equilibrium outcome in an infinitely repeated Bertrand duopoly by threats to revert forever to the competitive price if anyone deviates. For low discount factors, only the competitive price can be supported. Large number of firms in Bertrand model shrinks set of implicit agreements supportable via Folk Theorem by making limit on δ more stringent (MWG 405).

Implicit collusion in Cournot model (Kreps 524-526). Example12.AA.1 at MWG 422-423: Supporting zero payoffs in infinitely repeated Cournot duopoly (strategies yield zero-profit quantities followed forever by monopoly output until someone deviates).

Intellectual problems: Folk Theorem sets are independent, except perhaps for boundaries, of institutional structure, so theory provides little or no guide to modeling its effects. E.g. there's little difference in the Folk Theorem sets for repeated Cournot, Bertrand, and Stackelberg games (since each firm can unilaterally "blow" up the market by enforcing a zero price; note that definition of minimax payoff doesn't impose equilibrium or subgame-perfectness), and the Folk Theorem set for repeated Stackelberg is independent of which firm is the leader. Can't do comparative statics without theory of equilibrium selection.

Reputation in infinitely repeated interactions for product quality (one-sided Prisoner's Dilemma (Kreps 531-534)).

Entry deterrence with long-lived incumbent and short-lived entrants (Kreps 535).

Green and Porter (1984 *EMT*) and noise in implicit contracts (Kreps 515-523, 526-529); need to punish bad outcomes even though cheating never occurs in equilibrium. Effect of noise in limiting benefits of collusion, moderating optimal severity and/or duration of punishments.

Problems 9.B.12-13 at MWG 303, problems 12.D.1-5 and 12.AA2 at MWG 433-435; and problem 1 at Kreps 546-547

5. Games of Incomplete ("Asymmetric") Information

MWG 253-257 and 282-296; Kreps 425-437 and 463-489; Varian 279-282

Harsanyi's move by Nature trick and the generality of describing informational differences in games of *incomplete information* by *types* that parameterize preferences, drawn from common knowledge joint distribution (no need for beliefs about beliefs, etc.). *Harsanyi doctrine/common prior assumption*: any differences in players' beliefs can be viewed as derived from Bayesian updating of a common prior within a common model.

Strategy in game of incomplete information as complete type-contingent plan. Players need to make conjectures about alternative types, even their own and even though only one type is realized per player.

A pure-strategy *Bayesian Nash equilibrium* is a profile of decision rules (mapping each of each player's possible types into a strategy) that are in equilibrium in the game of complete information that arises before Nature chooses their types (Dn. 8.E.1 at MWG 255).

A profile of decision rules is a *Bayesian Nash equilibrium* if and only if for all types that have positive prior probability, a player's decision rule maximizes his expected payoff given his type, where the expectation is taken over other players' types conditional on his own type (Prop. 8.E.1 at MWG 255-256).

Thus, ex ante/complete information view of a game of incomplete information is equivalent to the *interim*/incomplete information view, in which players choose strategies after observing their types.

Recall definitions of beliefs, sequential rationality, *weak perfect Bayesian equilibrium*, and *sequential equilibrium* (MWG 282-296, Kreps 425-437).

Contracting example with Proposer's (R's) preferences u(y)=2, u(x)=1, u(z)=0, two types of Responder, C_1 with probability p and C_2 with probability (1-p), with C_1's preferences $v_1(x)=2$, $v_1(z)=1$, $v_1(y)=0$ and C_2's preferences $v_2(x)=2$, $v_2(y)=1$, $v_2(z)=0$, so C_1 won't accept y, R's favorite contract, and there is a tension between R's preferences and the probability of acceptance. Only C observes his type, but p and the rest of the structure are common knowledge. R's pure strategies are x and y, but C's now map his type and R's proposal into an accept or reject decisions, so he has 2x2x2x2=16 pure strategies, 4 chosen independently for each type. Draw extensive form with move by Nature first, then two decision nodes for R in the same information set, then 4 decision nodes for C, each in its own information set. There is a unique (unless p = ½) weak perfect Bayesian equilibrium, in which R proposes y if p<½, x if p>½, and either if p=½; C_1 accepts x but rejects y; and C_2 accepts x and y. Analysis easy because the only privately informed player has passive role.

Matching Pennies with random, independent payoff perturbations. *Purification* of mixed-strategy equilibrium by continuously distributed private information (MWG 257, Kreps 410).

Ultimatum Game with privately observed, continuously distributed outside option payoff. R proposes a division x for R, 1-x for C, and C accepts or rejects. If C accepts, R and C get payoffs x and 1-x. If C rejects, R gets payoff 0 and C gets payoff y with c.d.f. F(y), where F(0)=0, F(1)=1, and F(·) is continuously differentiable with positive density everywhere in between (e.g. uniform, with F(y)≡y when yϵ[0,1]). Any proposal risks rejection, with probability increasing in x. For most F(·) there is an essentially unique weak perfect Bayesian equilibrium, in which C accepts iff 1-x ≥ y (≥ rather than > without essential loss of generality, because the event 1-x = y has zero probability) and R proposes x*, 1-x*, where x* solves $\max_x$ xF(1-x). (When F(y)≡y, x*=½.) For some F(·) this problem has multiple solutions, in which case weak perfect Bayesian equilibrium is essentially nonunique. Analysis is again easy because the only privately informed player has a passive role.

Milgrom and Roberts' (1982 *EMT*) model of entry deterrence (Kreps 463-480, Figure 13.2 at 473). Two expected-profit maximizing firms, Incumbent and potential Entrant, choose Quantities, perfect substitutes, I in both of two periods, E only in second period. I has two possible unit costs, constant across periods, which only it observes: \$3 with probability ρ and \$1 with probability 1- ρ. E's unit cost is certain to be \$3. Both have fixed costs of \$3. ρ and the rest of the structure are common knowledge. In the first period, I observes its unit cost c and chooses Q, which determines P = 9 - Q. In the second period, E observes the first-period P and chooses whether or not to enter. If E enters, I and E are Cournot competitors in second period, taking into account whatever information is revealed in equilibrium by I's first-period P; if not, I is a monopolist in second period.

The analysis is hard because privately informed I plays an active role. I's first-period actions can signal its type to E, and in equilibrium both I and E must weigh the indirect, informational payoff implications of I's first-period decisions against their direct effects.

First analyze the Cournot subgame following entry, given E's beliefs (Kreps 475). If E assesses that c= 3 has probability μ, the Cournot equilibrium is Q_E = 2(2+μ)/3, Q_I|(c =1) = (10- μ)/3, Q_I|(c =3) = (7- μ)/3, with π_E = $4(2+\mu)^2/9$, not including its fixed cost of 3. Thus E enters iff $4(2+\mu)^2/9$ > 3, or μ > 0.598. E.g., if E knows c = 3, I and E each set Q_i = 2 and get π_i = 1 (=4-3), so it's profitable to enter. If E knows c = 1, I sets Q_I = 10/3 and E sets Q_E = 4/3 and gets π_E = -11/9, so it's not profitable to enter.

Now consider I's first-period decision. The first-period monopoly optimum is Q = 4, P = 5, π = 13 if c = 1; Q = 3, P = 6, π = 6 if c = 3. However, there is no weak perfect Bayesian equilibrium in which each type of I chooses its monopoly optimum in the first period. In such an equilibrium, E could infer I's type by observing P, and would enter if P = 6, thinking that c = 3. But then the high-cost type of I would get π = 6 in the first period and π = 1 in the second, less over the two periods than the π = 5 and π = 6 it could get (in the hypothesized equilibrium) by switching to P = 5 and thereby preventing E from entering.

The conclusion that there is no equilibrium of this kind does not depend on zero-probability inferences, and therefore holds for any stronger notion as well as weak perfect Bayesian equilibrium. Only one type needs to want to defect to break the equilibrium, and this is enough to invalidate it as a prediction even if that type is not realized. (Why?)

Now consider whether there can be a weak perfect Bayesian *pooling* equilibrium, in which both types of I charge the same price with probability one, and are therefore not distinguishable in equilibrium? (Looking for each possible kind of equilibrium is a characteristic form of analysis.)

If ρ < 0.598, there is a sequential (and weak perfect Bayesian) equilibrium in which: (i) each type of I sets P = 5 in the first period; (ii) E sticks with its prior ρ < 0.598 and therefore stays out if P ≤ 5 (in any weak perfect Bayesian pooling equilibrium, E *must* stick with its prior on the equilibrium path); (iii) E infers that I's costs are high and enters if P > 5; and (iv) entry leads to the Cournot equilibrium with E believing (as common knowledge) that I's costs are high. In this pooling equilibrium, the high-cost I successfully "hides behind" the low-cost I by giving up some profit in the first period to mimic the low-cost I, and both types of I successfully forestall entry.

To see that these strategies and beliefs are consistent with sequential equilibrium, note that: (i) E's strategy is sequentially rational, given its beliefs; (ii) the beliefs are consistent with Bayes' Rule on the equilibrium path; (iii) when c = 1, I charges its favorite first-period price and prevents entry, the best of all possible worlds; and (iv) when c = 3, the only way I could do better is by raising P above 5, but this would cause E to enter and thereby lower total profits. (Assuming the most pessimistic conjectures about consequences of deviations from equilibrium is a characteristic form of analysis, and yields largest possible set of weak perfect Bayesian equilibria.) Note that the beliefs also satisfy a natural *monotonicity* restriction, in that a higher P never lowers E's estimate that I's costs are high.

If ρ > 0.598, there is no weak perfect Bayesian pooling equilibrium. Such an equilibrium would always lead to entry, making the high-cost I unwilling to charge other than its first-period optimal monopoly price. The low-cost I would prefer a different price, even if it didn't prevent entry.

However, if ρ > 0.598 (or for any ρ) there is a *separating (screening, sorting)* sequential (hence weak perfect Bayesian) equilibrium in which: (i) the high-cost I charges its optimal monopoly price, 6, in the first period; (ii) the low-cost I charges 3.76 in the first period; (iii) E infers that costs are high if P>3.76 and therefore enters; (iv) E infers that costs are low if P≤3.76 and therefore stays out; (v) both types of I charge their monopoly price in the second period if there is no entry; and (vi) entry leads to the Cournot equilibrium with E believing (as common knowledge) that I's costs are high. In this separating equilibrium, the low-cost I successfully distinguishes itself from the high-cost I by distorting its first-period price enough to prevent the high-cost I from mimicking it. Entry occurs exactly when it would with complete information, and the only effect of incomplete information is the distortion of the low-cost I's first-period price, which benefits consumers and hurts the low-cost I. That the presence of alternative "bad" types hurts "good" types is typical.

To see that these strategies and beliefs are consistent with sequential equilibrium, note that: (i) E's strategy is again sequentially rational, given the hypothesized beliefs; (ii) the beliefs are (trivially) consistent with Bayes' Rule on the equilibrium path (and again monotonic); (iii) the low-cost I would like to set $P>3.76$ in the first-period, but that would lead to entry and reduce total profits (easy to check); and (iv) the high-cost I gets $\pi=6$ in the first period and $\pi=1$ following entry in the second, just above what it would get by setting $P\leq 3.76$ and forestalling entry (3.76 was chosen to make it just too costly for the high-cost I to mimic the low-cost I in this equilibrium). These arguments don't depend on ρ, so this profile is a weak perfect Bayesian equilibrium for any ρ.

Irrationality in finitely repeated Prisoner's Dilemma (summarize, refer to Kreps 480-489, 536-543).

Problems 8.E.1-3 at MWG 265, problems 9.C.1, 3-4, and 7 at MWG 304-305; and problems 17-18 at Kreps 457 and 2-4 at Kreps 498-501.

6. Adverse Selection MWG 436-448; Kreps 625-629; Varian 466-469
George Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* (August 1970); reading 15 in DR

Informational asymmetries and adverse selection in competitive labor market with many identical, risk-neutral, expected profit-maximizing firms, many workers with privately observed ability θ (= output) distributed on a compact interval with c.d.f. $F(\theta)$ and reservation ("home") wage $r(\theta)$. Full-information, efficient benchmark outcome has each worker working iff $r(\theta)\leq\theta$, each paid his θ. Inefficient equilibrium with $r(\theta)\equiv r$: Equilibrium wage $w^*=E\theta$ for those who accept. If $w\geq r$ all workers accept employment; if $w<r$ none do; which occurs determined by proportions of high- and low-ability workers. If too many low-productivity workers, firms unwilling to pay wage that any workers will accept, and there is too little employment; if too many high-productivity workers, firms pay wage that all accept, and there is too much employment in equilibrium. Asymmetric information prevents market from allocating workers efficiently between work and home.

When $r(\theta)$ varies with θ, can get market failure via *adverse selection* (informed agents' decisions hurt uninformed agents). Suppose $r(\theta)<\theta$ for all θ, so that all workers "should" work; that $r(\theta)$ is strictly increasing; and that there is a density of abilities θ. Then at any given wage, only the less able workers (those with $r(\theta)\leq w$) will work, so that higher wage rates raise the average productivity of those who accept employment. The equilibrium wage $w^*=E[\theta|\ r(\theta)\leq w^*]$. w^* must be $r(\theta)$ for the highest θ to get the best workers to work, but in Figure 13.B.1 at MWG 441 firms can't break even at this level. Thus the best workers don't work in equilibrium, the equilibrium is inefficient, and the equilibrium wage equals the average productivity of those who do work.

In cases like Figure 13.B.2 (left) at MWG 442, adverse selection causes complete market failure: no one works, even though all "should." Equilibrium can be unique as in Figure 13.B.1 or multiple and Pareto-ranked as in Figure 13.B.2 (right), with the high-wage equilibrium better for all workers and no worse for firms, who all earn zero profits in any equilibrium ("coordination failure").

In a two-stage game where firms first simultaneously choose wages and workers then choose among firms, when there is a density of abilities θ, the highest-wage competitive equilibrium is the unique subgame-perfect equilibrium (unless $w^*=r(\theta)$ for the lowest θ, in which case there can be multiple subgame-perfect equilibria, but all pure-strategy subgame-perfect equilibria yield workers the same payoffs as the highest-wage competitive equilibrium). Firms break lower-wage equilibria by raising wage and attracting higher-productivity workers (Proposition 13.B.1 at MWG 443-444).

Can use notion of *(incentive-)constrained Pareto-efficient allocation* to ask if the market does as well as possible, given limited information, and think about welfare effects of market intervention by planner who faces the same informational limitations as agents in the market. Planner must pay the same wage to all employed workers, and possibly different wage to all unemployed workers. He can implement the competitive outcome by setting $w=w^*$, and can enforce the highest-wage equilibrium. But he can do no better than that in this model because the high-wage equilibrium is constrained Pareto-efficient (Proposition 13.B.2 at MWG 447-448; the proof shows that any change from w^* hurts either low- or high-ability workers, and also addresses boundary issues).

Problems 13.B.1-9 at MWG 473-474 and problem 1 at Kreps 654.

7. Signaling and Screening

MWG 450-467; Kreps 629-652; Varian 469-471
Michael Spence, "Job Market Signalling," *Quarterly Journal of Economics* (August 1973); reading 18 in DR
Michael Rothschild and Joseph Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics* (November 1976); reading 17 in DR

Given inefficiency of competitive outcomes with asymmetric information, agents may consider various tactics to improve outcome (not that the agents care about efficiency, but inefficiency gives them opportunity to change the outcome). We consider two tactics, *signaling* (actions by informed agents to distinguish types) and *screening* (such actions by uninformed agents), which can be individually advantageous responses to asymmetric information, but need not increase efficiency. (*Sorting*, *separating*, and *pooling* tend to refer to equilibrium outcomes, while *signaling* and *screening* tend to refer to agents' actions.) I discuss both in a labor market example, following MWG 450-467 (and Kreps 629-637 and 645-649), though the original screening analyses of Stiglitz (monopoly) and Rothschild-Stiglitz (competition) were in an insurance market.

Spence's signaling model (MWG 450-460, Kreps 629-637; MWG assume unproductive education and Kreps assumes productive education, but the difference is inessential; I follow MWG): Two firms, one worker (can generalize). Market structure as in Section 6, but worker now has only two ability "types," with productivities $\theta_H>\theta_L>0$, where $0<\text{Prob}\{\theta=\theta_H\}=\lambda<1$. Only workers observe their types, but everyone knows λ, as common knowledge. $r(\theta_H)=r(\theta_L)=0$, so the unique equilibrium when workers can't signal has all workers employed at wage $w^*=E\theta$ and is Pareto-efficient. But workers can now choose education level e, continuously variable within a bounded interval, with differentiable and uniformly higher marginal ("single crossing property") and total costs for θ_L. Education has no effect on productivity (can relax). Although firms cannot directly observe ability, they can observe education levels, which in equilibrium might indirectly signal workers' abilities to firms because of the different costs of education for high- and low-ability workers.

The "rules" are as follows (Figure 13.C.1 at MWG 451 gives extensive form): (i) nature chooses the worker's type θ; (ii) worker observes type and chooses education level e; (iii) firms observe e and simultaneously make wage offers w_i; (iv) worker observes w_i and chooses between firms.

Add to weak perfect Bayesian equilibrium the condition that for all e (not just those chosen in equilibrium), both firms use a common posterior μ(e) to update their beliefs about the worker's ability and to predict each other's equilibrium w_i from the equilibrium offer functions. This added consistency of beliefs and strategies off the equilibrium path yields *perfect Bayesian equilibrium* (*PBE*), equivalent here to sequential equilibrium. A set of strategies and beliefs μ(e) is a PBE iff:

(i) the worker's strategy is optimal given the firms' strategies;

(ii) μ(e) is derived from the worker's strategy using Bayes' Rule whenever possible; and

(iii) the firms' wage offers following each possible e are in Nash equilibrium in the simultaneous-move wage offer game when the probability that the worker is of high ability is μ(e).

Just as in Bertrand duopoly, if the firms observe e and have beliefs μ(e), their unique equilibrium offers are both $\mu(e)\theta_H + (1- \mu(e))\theta_L$; the worker picks either of them, it doesn't matter which.

Lemma 13.C.1 at MWG 453: In any separating PBE, each worker type is paid its productivity.

Lemma 13.C.2 at MWG 454: In any separating PBE, a low-ability worker sets e = 0 (special to unproductive education, of course). Proof: e > 0 costs more, can't help because worker is separated.

Figures 13.C.5-6 at MWG 454-455 show separating PBEs with different supporting wage functions w*(e), each derived from common beliefs (hence between dotted lines). A high-ability worker chooses ê, the lowest e that a low-ability worker won't wish to imitate. Firms and workers behave optimally on and off equilibrium path: firms bid correctly and consistently for each e, and each worker type has a generalized tangency between its indifference curve and (w,e) opportunity locus.

Figure 13.C.7 at MWG 455 shows a separating PBE in which high-ability worker chooses e higher than ê, the minimum needed to separate from low-ability worker; its e can go all the way up to the e_1 that makes high-ability worker willing to mimic low-ability worker. These separating equilibria are Pareto-ranked; lowest-e is best. Low-ability workers do worse than when education is impossible. High-ability workers can do better or worse (Figure 13.C.8 at MWG 456; can do worse because can't duplicate no-education outcome, in which they were pooled with low-ability workers, and the education needed to separate is costly). The set of separating equilibria is independent of λ.

Figures 13.C.9-10 at MWG 457 show limits of pooling PBEs, with e anything from 0 to e', the level of e that makes low-ability worker indifferent between being identified as low-ability at e = 0 and being pooled at e = e'. Firms and workers both behave optimally on and off equilibrium path: firms because they bid correctly and consistently for each e, and worker types because each has a generalized tangency between its indifference curve and (w,e) opportunity locus. Pooling equilibria are again Pareto-ranked, with the e = 0 one best for both worker types and the firms not caring. All pooling equilibria are weakly Pareto-dominated by the no-education-is-possible equilibrium.

Figure 13.C.7 illustrates the use of equilibrium refinements to break separating equilibria in which high-ability worker chooses e higher than ê. An e between ê and e_1 is *equilibrium-dominated* for the low-ability type, in that it is dominated if we assume equilibrium beliefs and bids by firms. This kind of argument, which goes beyond sequential equilibrium (and monotonicity, etc.) to restrict out-of-equilibrium beliefs, is called in its simplest form the *intuitive criterion.* In this model one can use such arguments to rule out separating equilibria with high-ability workers setting e "too high" and all pooling equilibria (Figures 13.C.9-10). The result is a unique prediction of the *outcome*, that of the separating equilibria with different beliefs in Figures 13.C.5-6. This outcome may not be constrained Pareto-efficient. If the no-signaling equilibrium Pareto-dominates the separating equilibrium, banning signaling is a Pareto-improvement. If not, market intervention by setting separate wages for workers with e above and below a properly chosen cutoff (Figure 13.C.11 at MWG 458) may still allow a Pareto-improvement by making both worker types better off while allowing firms to break even by cross-subsidization (losing on low-ability workers but gaining on high-ability workers). In more realistic models, educational signaling can improve matching between workers and jobs and/or enhance productivity. However, the desire to separate can still lead to excessive education relative to what would be optimal with observable ability.

Rothschild-Stiglitz competitive screening model in the labor-market example (MWG 460-467, Kreps 638-645 and 649-650; the Stiglitz monopoly screening model is a special case of the agency models in Section 8, discussed at MWG 500-501 and Kreps 661-680):

Market structure is almost the same as in Section 7: Two firms, but many workers (inessential). Workers have two ability "types" with productivities $\theta_H > \theta_L > 0$, where $0 < \text{Prob}\{\theta = \theta_H\} = \lambda < 1$. $r(\theta_H) = r(\theta_L) = 0$. Only workers observe their types, but λ is common knowledge. Workers no longer choose education, but firms can offer contracts with different "task levels" (e.g. hours) to induce workers to reveal their types by their choice of contract. Task level has no effect on productivity (can relax). A type-θ worker with wage w and task level $t \geq 0$ has utility $u(w,t|\theta) = w - c(t,\theta)$, where $c(0,\theta) = 0$, $c_t(t,\theta) > 0$, $c_{tt}(t,\theta) > 0$, $c_\theta(t,\theta) < 0$ for all $t > 0$, and $c_{t\theta}(t,\theta) < 0$ ("single crossing property").

The "rules" are as follows: (i) nature chooses the workers' types, θ; (ii) firms simultaneously offer sets of (any desired finite number of, but two is enough) contracts, each of which is a pair (w,t); (iii) workers observe their types and each type chooses one of the offered contracts or no contract (assume for simplicity that workers who are indifferent between contracts choose the one with lower t, workers who are indifferent between a contract and no contract choose the contract, and workers whose most preferred contract is offered by both firms choose each with probability ½).

Study pure-strategy subgame-perfect Nash equilibria (SPNE) throughout. Suppose first that workers' types are observable, so firms can condition offers on a worker's type, offering a contract (w_L,t_L) restricted to low-ability workers and a contract (w_H,t_H) restricted to high-ability workers.

Proposition 13.D.1 at MWG 461-462 (Figure 13.D.1): In any SPNE of the game with observable worker types, a worker of type θ_i accepts contract $(w_i^*, t_i^*) = (\theta_i, 0)$, and firms earn zero profits. Proof: Firms gain by replacing inefficient contracts with $t_i > 0$ by $t_i = 0$, and competition drives w_i up to θ_i.

Now suppose that workers' types are unobservable, so that any offered contract can be accepted by a worker of either type. The full-information outcome of Proposition 13.D.1 is no longer attainable, because low-ability workers prefer the high-ability contract to the low-ability contract, and they can no longer be prevented from accepting it. We will look for pooling or separating equilibria.

Lemma 13.D.1 at MWG 462-463: In any equilibrium, pooling or separating, both firms earn zero profits. Proof: If (w_L,t_L) and (w_H,t_H) are the contracts chosen by low- and high-ability workers, respectively, and firms have positive total profits π, at least one firm must make profit≤π/2. Such a firm can attract all low-ability workers by offering $(w_L+\varepsilon,t_L)$ and all high-ability workers by offering $(w_H+\varepsilon,t_H)$ for some small $\varepsilon > 0$. Since ε can be as small as desired, that firm can gets profits as close as desired to π, and therefore has a profitable deviation unless $\pi \leq 0$. But π can't be negative because firms aren't required to offer any contracts, so in equilibrium $\pi = 0$.

Lemma 13.D.2 at MWG 463: No pooling equilibria exist. Proof: If there is a pooling equilibrium contract (w^p,t^p), by Lemma 13.D.1 it lies on the pooled break-even (0-profit) line in Figure 13.D.3 at MWG 463. Then either firm can gain by deviating to a contract in the shaded lens with w< θ_H, which attracts all high-ability workers and no low-ability workers and thus yields positive profits.

Lemma 13.D.3 at MWG 463: If (w_L,t_L) and (w_H,t_H) are contracts chosen by low- and high-ability workers in a separating equilibrium, then w_L=θ_L and w_H=θ_H so both yield zero profits. Proof: If w_L<θ_L either firm could get positive profits by offering only a contract with w a little above w_L, which all low-ability workers would accept, and which would be profitable for both low- and high-ability workers. This contradicts Lemma 13.D.1, so w_L≥θ_L in any separating equilibrium. If w_H<θ_H as in Figure 13.D.4 at MWG 463, then (w_L,t_L) must lie in the lens in Figure 13.D.4 above the w_L=θ_L line as shown, by self-selection and Lemma 13.D.1 (0 profits). But then either firm could get positive profits by deviating to a contract in the upper lens below w_H=θ_H line, like (w~,t~) in Figure 13.D.4, which would attract all high-ability workers. Thus w_H≥θ_H in any separating equilibrium. Since firms break even in any equilibrium by Lemma 13.D.1, in fact w_L=θ_L and w_H=θ_H.

Lemma 13.D.4 at MWG 464: In any separating equilibrium, low-ability workers accept (θ_L,0), the same contract they would receive in a full-information competitive equilibrium. Proof: By Lemma 13.D.3, w_L=θ_L in any separating equilibrium. If t_L>0 in such an equilibrium, a firm could do better by offering a contract with lower w_L and t_L, attracting all low-ability workers as in Figure 13.D.5.

Lemma 13.D.5 at MWG 464: In any separating equilibrium, high-ability workers accept (θ_H,$t_H^{\wedge}$), where $t_H^{\wedge}$ is chosen so low-ability workers are indifferent between (θ_L,0) and (θ_H,$t_H^{\wedge}$) as in Figure 13.D.6 at MWG 464, so that θ_H–c($t_H^{\wedge}$, θ_L)=θ_L–c(0, θ_L). Proof: By Lemmata 13.D.3-4, w_H=θ_H and (w_L,t_L)=(θ_L,0). For low-ability workers to accept (θ_L,0), t_H≥$t_H^{\wedge}$ in Figure 13.D.6. If the high-ability contract (θ_H,t_H) has t_H>$t_H^{\wedge}$ as in Figure 13.D.6, then either firm can get positive profits by offering an *additional* contract with lower w_H and t_H as shown in Figure 13.D.6, which attracts all of the high-ability workers and does not change the choices of low-ability workers. Thus, in any separating equilibrium, the high-ability contract must be (θ_H,$t_H^{\wedge}$).

Proposition 13.D.2: In any subgame-perfect equilibrium of the screening game, low-ability workers accept contract (θ_L,0) and high-ability workers accept contract (θ_H,$t_H^{\wedge}$) in Figure 13.D.6, where θ_H–c($t_H^{\wedge}$, θ_L)=θ_L–c(0, θ_L).

Proposition 13.D.2 tells what a separating equilibrium must look like if one exists, but does not tell us that such an equilibrium exists. Consider the candidates for a separating equilibrium in Figures 13.D.7(a-b). By construction, for any λ, neither firm can gain from deviating in a way that attracts either all high- or all low-ability workers. However, varying λ allows us to move Eθ anywhere between θ_H and θ_L without affecting the candidate separating equilibrium. And in Figure 13.D.7(b) (but not (a)), Eθ is high enough that a firm can gain by deviating to a contract such as (w~,t~) that attracts all workers to a single *pooling* contract. In this case, since no pooling equilibrium ever exists, no equilibrium of any kind exists (in pure strategies; equilibrium does exist in mixed strategies, but the interpretation of mixed-strategy equilibria in this model is problematic).

As in the signaling model's best separating equilibrium, screening equilibria are Pareto-inefficient, and low-ability workers are worse off than when screening is impossible. However, when a screening equilibrium exists it must make high-ability workers better off (whenever screening would hurt high-ability workers, a pooling contract breaks the screening equilibrium candidate). When they exist, screening equilibria are (with a qualification) constrained Pareto-efficient.

Problems 13.C.1-6 and 13.D.1-4 at MWG 474-476; and problems 2-5, 7-10 at Kreps 654-660

8. Agency

MWG 477-506; Kreps 577-614 and 661-674; Varian 441-466; McMillan 91-129
Oliver Hart and Bengt Holmstrom, "The Theory of Contracts," in Truman Bewley, editor, *Advances in Economic Theory, Fifth World Congress*, Cambridge 1987

Consider a relationship between two people: a *principal* (sometimes called "owner" in MWG) who could benefit from delegating a decision that affects his welfare to an *agent* (sometimes called "manager" in MWG) who has relevant skills or private information. The agent has different preferences over decisions than the principal would if he were fully informed, and the principal cannot control the agent's decisions (because he cannot observe them, or for other, unmodeled, reasons). But the principal can design a contract or incentive scheme to influence agent's decisions.

Distinction between *hidden actions/moral hazard* (e.g. fire prevention, manager's effort choice that influences owner's profit, borrower's investment decisions that influence lender's return on loan) *and hidden information/adverse selection* (e.g. insurer unable to observe consumer's risk class). Distinction is independent of signaling-screening distinction. Applications often have some of both. Analytically similar in that in each case the principal cannot observe the agent's *decision rule*.

Hidden-action analysis: Agent chooses one-dimensional effort level e from set E, which is costly to the agent. e influences the principal's profit π. Principal wishes to maximize Eπ net of what he pays agent, but cannot observe (or directly control) e. If relationship between e and π were deterministic, invertible, principal could infer e from π, and thereby control e; so assume π has a conditional density f(π|e)>0 for all eεE and all $\pi \varepsilon [\underline{\pi}, \bar{\pi}]$, making any value of π consistent with any value of e.

Special case with two effort levels, e_H and e_L: $f(\pi|e_H)$ first-order stochastically dominates $f(\pi|e_L)$ (i.e. $F(\pi|e_H) \leq F(\pi|e_L)$) for all $\pi \varepsilon [\underline{\pi}, \bar{\pi}]$, with strict inequality for nonnegligible set of π's. $E_{F(\pi|eH)}\pi > E_{F(\pi|eL)}\pi$, so principal prefers agent to choose e_H, other things equal. Agent chooses $e \varepsilon \{e_H, e_L\}$ to maximize $E[v(w)-g(e)]$, w is wage, v(·) strictly increasing and weakly concave so agent risk-averse in income, and $g(e_H)>g(e_L)$ so agent dislikes effort. Principal is risk-neutral and maximizes $E[\pi\text{-}w]$.

Ultimatum model of contracting process (standard in principal-agent literature): Principal proposes contract to agent, which agent can accept or reject. Acceptance yields binding contract, rejection yields agent reservation utility $\underline{u}$, a proxy for agent's best alternative in the market, assumed exogenous. Assume subgame-perfect equilibrium (SPNE) throughout.

Proposition 14.B.1 at MWG 480-481: When the agent's effort is observable, an optimal (uniquely when $v(\cdot)$ is strictly concave) contract for the principal specifies that the agent choose the effort e^* that solves $\max_{e\varepsilon\{eH,eL\}} [E_{F(\pi|e)}\pi - v^{-1}(\underline{u}+g(e))]$ and pays the agent a fixed wage $w^*= v^{-1}(\underline{u}+g(e^*))$.

Proof: When the agent's effort is observable, a contract specifies agent's effort $e\varepsilon\{e_H,e_L\}$ and wage $w(\pi)$. The principal's problem is $\max_{e\varepsilon\{eH,eL\},w(\pi)} E_{F(\pi|e)}[\pi - w(\pi)]$ s.t. $E_{F(\pi|e)}v(w(\pi)) - g(e)\geq\underline{u}$ (*participation* or *individual rationality* constraint). First consider the best $w(\pi)$ given e, which solves $\min_{w(\pi)}E_{F(\pi|e)}\, w(\pi)$ s.t. $E_{F(\pi|e)}v(w(\pi)) - g(e)\geq\underline{u}$. Constraint always binding, with Lagrange multiplier γ and first-order condition $\gamma = 1/v'(w(\pi))$ for all π. Given e, if $v(\cdot)$ is strictly concave this implies that $w(\pi)=w^*(e)$ for all π, and if $v(\cdot)$ is weakly concave $w(\pi)=w^*(e)$ is still one optimum. (The best way for a risk-neutral principal to get a risk-averse agent up to utility level $\underline{u}$ is for the principal to bear all the risk.) Thus $v(w^*(e)) - g(e)=\underline{u}$, so $w^*(e)=v^{-1}(\underline{u}+g(e))$, with $w^*(e)$ increasing in e. Given $w^*(e)=v^{-1}(\underline{u}+g(e))$, the best e, e^*, solves $\max_{e\varepsilon\{eH,eL\}} [E_{F(\pi|e)}\pi - v^{-1}(\underline{u}+g(e))]$.

Proposition 14.B.2 at MWG 482-483: When the agent's effort is *un*observable but the agent is risk-neutral, the optimal contract leads to the same effort and expected utilities for principal and agent as when effort is observable.

Proof: Suppose the principal sets $w(\pi)\equiv\pi-\alpha$ for some constant α ("selling the project (for α) to the agent"). The agent then chooses e to solve $\max_{e\varepsilon\{eH,eL\}}[E_{F(\pi|e)}w(\pi) - g(e)] = E_{F(\pi|e)}\pi - \alpha - g(e)$. When $v(w)\equiv w$, $v^{-1}(w)\equiv w$, so this problem has the same solution e^* that solves $\max_{e\varepsilon\{eH,eL\}} [E_{F(\pi|e)}\pi - v^{-1}(\underline{u}+g(e))]$ in Proposition 14.B.1 (the maximands differ by a constant). Setting $\alpha=\alpha^*$ where $E_{F(\pi|e^*)}\pi - \alpha^* - g(e^*)=\underline{u}$ satisfies the agent's participation constraint and yields the principal utility $\alpha^*=E_{F(\pi|e^*)}\pi - g(e^*) - \underline{u}$, the same as his utility in the optimal contract with observable effort in Proposition 14.B.1. The optimal contract with unobservable effort could not possibly improve on this.

When the agent's effort is *un*observable and the agent is risk-averse, there is a tension between efficient risk-sharing and providing efficient incentives for the agent that makes the problem nontrivial. E.g. perfect fire insurance dilutes incentives to take care against fire. Optimal contract is a second-best compromise.

Proposition 14.B.3 at MWG 483-488: When the agent's effort is *un*observable, the agent is risk-averse, and there are two possible effort choices, the optimal compensation scheme for implementing e_H satisfies $1/v'(w(\pi)) = \gamma + \mu[1 - f(\pi|e_L)/f(\pi|e_H)]$, gives the agent expected utility $\underline{u}$, and involves a larger Ew* than when effort is observable. The optimal compensation scheme for implementing e_L involves the same fixed w as if e were observable.

Whenever the optimal effort with observable e would be e_H, the nonobservability of e causes a welfare loss: Either it is still optimal to implement e_H, in which case the agent faces avoidable risk which the principal must compensate him for (the agent still gets $\underline{u}$); or it is now too expensive to implement e_H, and the principal implements e_L even though e_H would allow a Pareto-improvement. (The fact that unobservable e makes incentive constraints bind and distorts effort downward may not be true for more than two effort levels (MWG 502-504, Exercise 14.B.4 at MWG 507).)

Proof: When e is unobservable the principal's optimal contract specifies a wage w(π). The best w(π) given e solves $\min_{w(\pi)} E_{F(\pi|e)} w(\pi)$ s.t. (i) $E_{F(\pi|e)} v(w(\pi)) - g(e) \geq \underline{u}$ (*participation* or *individual rationality* constraint) and (ii) e solves $\max_{e'} E_{F(\pi|e')} v(w(\pi)) - g(e')$ (*incentive compatibility constraint*).

If it is desired to *implement* e_L, it is optimal for the principal to offer a fixed wage payment $w^*(e_L) = v^{-1}(\underline{u} + g(e_L))$, as if he were specifying e_L when effort is observable. This makes agent choose e_L, because effort doesn't affect w and he prefers e_L, other things equal, and yields agent $\underline{u}$ just as when effort is observable. Optimal contract with unobservable effort could not possibly improve on this.

If it is desired to implement e_H, constraint (ii) becomes $E_{F(\pi|eH)} v(w(\pi)) - g(e_H) \geq E_{F(\pi|eL)} v(w(\pi)) - g(e_L)$. Again letting $\gamma \geq 0$ and $\mu \geq 0$ be the Lagrange multipliers on constraints (i) and (ii) respectively, w(π) must satisfy the following first-order condition for all π:
$-f(\pi|e_H) + \gamma v'(w(\pi)) f(\pi|e_H) + \mu[f(\pi|e_H) - f(\pi|e_L)] v'(w(\pi)) = 0$, *or* $1/v'(w(\pi)) = \gamma + \mu[1 - f(\pi|e_L)/f(\pi|e_H)]$.

When $e = e_H$, both constraints bind, because the agent would like to set $e = e_L$:

Lemma 14.B.1 at MWG 484: In any solution to the principal's problem with $e = e_H$, $\gamma > 0$ and $\mu > 0$.

Proof: Because $f(\pi|e_H)$ first-order stochastically dominates $f(\pi|e_L)$, there must be an open set of π throughout which $f(\pi|e_L)/f(\pi|e_H) > 1$. But if $\gamma = 0$ and $\mu \geq 0$, this contradicts the first-order condition. And if $\mu = 0$, the first-order condition implies a fixed wage payment, which implements e_L, not e_H.

Given this, this first-order condition says that the agent gets a "base payment" (in utility) that is independent of π plus a "bonus" that is higher to the extent that π is evidence (in the sense of the likelihood ratio $f(\pi|e_L)/f(\pi|e_H)$) that he chose e_H. This evidence affects the bonus not because the principal doubts that the agent chose e_H; in equilibrium, the principal knows this. Paying the agent partly according to the evidence that he chose e_H is just the cheapest way to get him to choose e_H.

We can also use the first-order condition to ask if w(π) must be increasing. Surprisingly, this is not true without further restrictions on f(·), because even when $f(\pi|e_H)$ first-order stochastically dominates $f(\pi|e_L)$, $f(\pi|e_L)/f(\pi|e_H)$ need not be decreasing in π. The *monotone likelihood ratio property* says that $f(\pi|e_L)/f(\pi|e_H)$ is decreasing in π, so that higher π is evidence in favor of e_H (Fig. 14.B.1, MWG 485-486, and Kreps 494-495 give examples to show why this property is necessary).

One can also use the first-order condition to prove the Mirrlees-Holmstrom *Sufficient Statistic Theorem* (MWG 487-488): If (and only if) π is a sufficient statistic for the agent's choice of e, there is no gain to allowing w to depend on any other available indirect measure of e.

The first-order condition shows that the optimal incentive scheme is generally highly nonlinear and sensitive to the details of the environment, including the distribution f(·). By contrast, real-world incentive schemes (e.g. sharecropping), tend to be simple and *robust* to environmental details. Why this is true is still largely an open question; MWG 488 discuss a possible explanation. Inefficiency makes devices like monitoring and cross-checking useful. MWG 488 discuss extensions to multiple agents with relative performance evaluation (*tournaments*), long-term relationships, competition for agents among multiple principals, and multidimensional effort.

Refer to MWG 504 and Kreps 604-608 on the "first-order approach" (different from above use of first-order condition) as imperfect alternative to min w(π) given e when e is continuously variable. Two problems: failure of second-order conditions and discontinuities of e* in w(π).

Hidden-information analysis (MWG 488-501): Almost the same model as for hidden actions, but now the agent chooses $e\varepsilon[0,\infty)$, e is observable, and the agent's cost of effort is unobservable. The principal's gross profit (net of wage payments) is π(e), with π(0)=0, π'(e)>0, and π"(e)<0 for all e. The agent's reservation utility is $\underline{u}$, and the agent's vN-M utility function is $u(w,e,\theta)\equiv v(w-g(e,\theta))$ where v"(·)<0, so the agent is risk averse in income. g(·) measures the cost of effort, with $g(0,\theta)\equiv 0$ for all θ and, for all e>0, $g_e(\cdot)>0$, $g_{ee}(\cdot)>0$, $g_\theta(\cdot)<0$, and $g_{e\theta}(\cdot)<0$ (the "single-crossing property"), so that e has positive and increasing marginal cost, and both are decreasing in θ. Focus here on the special case with two possible θs, θ_H and θ_L, with commonly known probabilities $\lambda\varepsilon(0,1)$ and 1-λ.

Ultimatum contracting: Principal proposes a contract, which agent can accept or reject. Acceptance yields binding contract, rejection yields agent reservation utility $\underline{u}$, a proxy for his best alternative in the market, assumed exogenous. Assume subgame-perfect Nash equilibrium (SPNE) throughout.

Consider first the benchmark case where θ is observable, so that the principal can specify the effort level e_i and wage w_i *contingent* on each realization of θ, θ_i. In the two-outcome case the contract specifies two wage-effort pairs, (w_H,e_H) and (w_L ,e_L), and the principal chooses these to solve

$$\max_{(wH,eH)\geq 0,\ (wL ,eL) \geq 0} \lambda[\pi(e_H)-w_H] + (1-\lambda)[\pi(e_L)-w_L]$$

$$\text{s.t. } \lambda v(w_H-g(e_H,\theta_H)) + (1-\lambda)v(w_L-g(e_L,\theta_L))\geq \underline{u}$$

Proposition 14.C.1 at MWG 492: When θ is observable, the optimal contract for the principal involves an effort level e_i^* in state θ_i such that $\pi'(e_i^*)=g_e(e_i^*,\theta_i)$, and fully insures the agent, setting the wage in each state θ_i at the level w_i^* such that $v(w_i^*-g(e_i^*,\theta_i))=\underline{u}$. Proof: The participation constraint must bind at the solution, because otherwise the principal could increase his profit by lowering wages. Letting γ≥0 be the multiplier on this constraint, we have the first-order conditions:

(14.C.2) $$-\lambda + \gamma\lambda v'(w_H^*-g(e_H^*,\theta_H))=0$$

(14.C.3) $$-(1-\lambda) + \gamma(1-\lambda)v'(w_L^*-g(e_L^*,\theta_L))=0$$

(14.C.4) $$\lambda\pi'(e_H^*) - \gamma\lambda v'(w_H^*-g(e_H^*,\theta_H))g_e(e_H^*,\theta_H)\leq 0, \text{ and } =0 \text{ if } e_H^*>0$$

(14.C.5) $$(1-\lambda)\pi'(e_L^*) - \gamma(1-\lambda)v'(w_L^*-g(e_L^*,\theta_L))g_e(e_L^*,\theta_L)\leq 0, \text{ and } =0 \text{ if } e_L^*>0$$

Combining (14.C.2) and (14.C.3) yields the standard condition for efficient insurance of a risk-averse party (the agent) by a risk-neutral party (the principal).

(14.C.6) $$v'(w_H^*-g(e_H^*,\theta_H)) = v'(w_L^*-g(e_L^*,\theta_L)).$$

Because $v''(\cdot)<0$, (14.C.6) implies that $w_H^*-g(e_H^*,\theta_H))=w_L^*-g(e_L^*,\theta_L)$ and $v(w_H^*-g(e_H^*,\theta_H))= v(w_L^*-g(e_L^*,\theta_L))$, and because the participation constraint is binding, the agent has utility $\underline{u}$ in each state. Because $g_e(0,\theta)=0$ and $\pi'(0)>0$, (14.C.4) and (14.C.5) must both hold with equality and with $e_H^*>0$ and $e_L^*>0$. Combining (14.C.2) and (14.C.4) and (14.C.3) and (14.C.5) yields

(14.C.7) $$\pi'(e_i^*)=g_e(e_i^*,\theta_i),\ i=L,H,$$

the condition for efficient effort choice, requiring that the marginal benefit of effort equals its marginal (utility) cost in each state (Figure 14.C.1 at MWG 491). The principal's profit in state i is

$$\Pi_i^*= \pi(e_i^*) – v^{-1}(\underline{u}) - g(e_i^*,\theta_i),\ i=L,H.$$

From (14.C.7), $g_{e\theta}(e,\theta)<0$, $\pi''(e)<0$, and $g_{ee}(e,\theta)>0$ imply $e_H^*>e_L^*$ (Figure 14.C.2 at MWG 492). This completes the proof, showing that when θ is observable, a risk-neutral principal fully insures a risk-averse agent and specifies fully efficient effort for each realization of the state θ_i.

When θ is unobservable, the principal's optimal contract must balance the provision of insurance for the agent against the need to give the agent incentives to make e vary appropriately with θ. (e is observable but θ is unobservable, so the relationship between e and θ is unobservable.) The first-best outcome of Proposition 14.C.1 is no longer attainable, because the agent always prefers (w_L^*,e_L^*) to (w_H^*,e_H^*) (Figure 14.C.2). Thus if the agent is asked to report θ (directly, or indirectly by his choice of effort) he will always report $\theta=\theta_L$, and the principal will not realize the first-best outcome. In characterizing the optimal contract in this case, we must consider the agent's incentives to misrepreesent θ and how this affects the outcome. The task is simplified by the following general result, which shows that, in a sense, there is no loss of generality in restricting attention to contracts in which the agent is asked to report θ (a *direct revelation mechanism*), and for which truthful reporting is consistent with equilibrium (so the mechanism is *incentive-compatible*).

Proposition 14.C.2 at MWG 493 (Revelation Principle): In determining the optimal contract, the principal can without loss of generality restrict attention to contracts in which: (i) after the agent observes θ, he is required to report it; (ii) the contract specifies an outcome for each possible report; and (iii) for every possible realization of θ, the agent finds it optimal to report θ truthfully. Proof: Given a particular selection among any multiple equilibria that exist in the game following a set of contract proposals by the principal, one can collapse any contract that creates an incentive for the agent to lie into an equivalent contract that specifies the outcome that lying yields in equilibrium.

Now consider the case where θ is unobservable, under the simplifying assumption that the agent is infinitely risk averse (maximin or limit of finite risk aversion). Write the principal's problem as

$$\max_{(wH,eH)\geq 0,\ (wL\ ,eL)\geq 0} \lambda[\pi(e_H)-w_H] + (1-\lambda)[\pi(e_L)-w_L] \text{ s.t.}$$

(i) $w_L-g(e_L,\theta_L) \geq v^{-1}(\underline{u})$ ("participation" or "individual rationality" constraint for θ_L)

(ii) $w_H-g(e_H,\theta_H) \geq v^{-1}(\underline{u})$ ("participation" or "individual rationality" constraint for θ_H)

(iii) $w_H-g(e_H,\theta_H) \geq w_L-g(e_L,\theta_H)$ ("incentive-compatibility" or "self-selection" constraint for θ_H)

(iv) $w_L-g(e_L,\theta_L) \geq w_H-g(e_H,\theta_L)$ ("incentive-compatibility" or "self-selection" constraint for θ_L), where (w_L,e_L) and (w_H,e_H) are now interpreted as what happens when the agent announces θ_L or θ_H. (There is no loss of generality here, by the Revelation Principle.) The participation/individual rationality constraints are given for the "interim" case where the agent observes his type before contracting, but with an infinitely risk averse agent this formulation applies equally well to the "ex ante" case where the agent signs the contract before observing his type. The interim case is often more relevant, and in other models may have different implications than the ex ante case.

Lemma 14.C.1 at MWG 495 (Figure 14.C.3 at MWG 495): Constraint (ii) is never binding, and can be ignored. Proof: From (i) and (iii), w_H-g(e_H,θ_H) ≥ w_L-g(e_L,θ_H) ≥ (because g(e_L,θ_H)≤g(e_L,θ_L)) w_L-g(e_L,θ_L)≥v^{-1}($\underline{u}$). (Intuition: If Low agents are happy to sign up, High agents must be even happier.)

Lemma 14.C.2 at MWG 495-496: An optimal contract must have w_L-g(e_L,θ_L) = v^{-1}($\underline{u}$), so constraint (i) is binding. Proof: Otherwise the principal could reduce both w_H and w_L by ε>0, preserving incentive-compatibility and increasing profits. (Intuition: Since High agents are always happier than Low agents, and the principal can screw both in a balanced way that does not interfere with incentive-compatibility, it is optimal for the principal to screw Low agents to the wall ($\underline{u}$).)

Lemma 14.C.3 at MWG 496-497: In any optimal contract: (i) e_L≤e_L*; and (ii) e_H=e_H* (e_L* and e_H* are from the optimal contract with observable θ). Proof (Figures 14.C.4-6 at MWG 496-497):

(i): By Lemma 14.C.2 and incentive-compatibility, (w_L,e_L) must be on upper boundary of shaded region in Figure 14.C.4. If e_L≥e_L*, the principal can increase profit by sliding (w_L,e_L) down the agent's $\underline{u}$ indifference curve to (w_L*,e_L*) in Figure 14.C.5, leaving both Low and High agents' utilities unchanged and continuing to satisfy the incentive-compatibility constraints.

(ii): Given (w_L,e_L^) with e_L≤e_L* as in Figure 14.C.6, the principal must find the (w_H,e_H) in the shaded region in Figure 14.C.6 that maximizes his profit in state θ_H. The solution occurs at a tangency like that at (w_H~,e_H*) in the figure, with e_H=e_H* because the only binding constraint that involves both e_H and θ_H is (iii), the incentive-compatibility constraint for the High agent.

The fact that only the incentive-compatibility constraint for the High agent is binding is common to such analyses. (With more than two types, this property generalizes to: only incentive-compatibility constraints for adjacent types bind, and they only bind in the "downward" direction.)

Lemma 14.C.4 at MWG 497-498 (also see Appendix B at MWG 504-506): In any optimal contract, $e_L<e_L*$. Proof: Start with $(w_L,e_L)= (w_L*,e_L*)$ as in Figure 14.C.7 at MWG 497, which by Lemma 14.C.3 determines the state θ_H outcome, $(w_H\sim,e_H*)$ in the figure. Principal's overall expected profit with $(w_L,e_L)= (w_L*,e_L*)$ is a $(\lambda, 1-\lambda)$-weighted average of his profits in states θ_H and θ_L, which can be read off the vertical axis in the figure (because $\pi(0)=0$, the principal's profit = -w). Sliding (w_L,e_L) a small amount down the Low agent's indifference curve, to $(w_L\hat{},e_L\hat{})$ in Figure 14.C.8(a) at MWG 498 (note typo in label of $(w_L\hat{},e_L\hat{})$) yields a zeroth-order reduction in profit in state θ_L, because it involves a small change in (w_L,e_L) away from the first-best (w_L*,e_L*) in that state (Envelope Theorem). However, it relaxes the incentive-compatibility constraint in state θ_H and thereby allows the principal to lower w_H by a first-order amount (Figure 14.C.8(b)). On balance, this increases the principal's profit. The more likely is θ_H, the more the principal is willing to distort the θ_L outcome to get higher profits in θ_H. (Follows from Kuhn-Tucker conditions; see MWG App. B at 504-506.)

Proposition 14.C.3 at MWG 499-500: To sum up, in the hidden-information principal-agent model with an infinitely risk-averse agent, the optimal contract sets $e_H=e_H*$ and $e_L<e_L*$, and the agent is inefficiently insured, getting utility > $\underline{u}$ in state θ_H and utility $\underline{u}$ in state θ_L. The principal's expected profit is lower than when θ is observable, while the infinitely risk-averse agent's utility is the same.

The conclusions would be the same if π were not publicly observable, in which case we could allow θ to affect the relationship between π and e (replacing $\pi(e)$ by $\pi_L(e)$ and $\pi_H(e)$). We couldn't do this if π were observable, because then the principal could infer θ from π and the specified e.

Stiglitz's (1977 *Review of Economic Studies*) analysis of monopolistic screening with adverse selection is just like this, except that the principal's profit depends directly on the agent's private information (MWG 500-501). Although this makes little difference, it is instructive to give Stiglitz's analysis, without assuming an infinitely risk-averse agent. Here I follow Kreps 661-674.

Basic model (Figure 18.3 at Kreps 665): Risk-neutral, expected-profit maximizing insurer (principal), risk-averse consumer (agent) with probability π_i that endowment will be Y_2 and probability $1- \pi_i$ that endowment will be Y_1, i=H,L, where $\pi_H>\pi_L$. Ultimatum contracting as before.

Graphing consumer's indifference curve in (y_1,y_2)-space, slope = $-(1-\pi)/\pi$ along 45˚ line in Figure 18.3. Risk-neutral insurer's indifference curves are linear with slope $-(1-\pi)/\pi$. Efficient insurance thus involves tangency on the 45˚ line, so risk-averse consumer is perfectly insured. $Y_2<Y_1$, so Y_2 is the "accident" outcome and the consumer's endowment (Y_1,Y_2) is below the 45˚ line. Thus, optimal monopolistic contract when insurer knows π_i is on the 45˚ line where it intersects the consumer's indifference curve through his endowment (Figure 18.3 at Kreps 665). (Competition with known π_i would also yield a contract on the 45˚ line, but drive insurers' expected profits to 0.)

When insurer doesn't know π_i but does know prior, ρ, that i=H, would like to use the full-information-optimal monopolistic contracts just derived. But because low-risk consumer cares less about y_2, his indifference curves are uniformly steeper in (y_1,y_2)-space than a high-risk consumer's (Figure 18.2 at Kreps 664). The optimal monopolistic contract for low-risk consumers will then look better for high- as well as low-risk consumers, so the insurer won't get the anticipated level of profit by offering the full-information-optimal monopolistic contracts.

When π_i is unknown, we can characterize the optimal contracts as follows. No loss of generality in restricting insurer to two contracts (two=number of consumer types), one designed for high-risk consumers and one for low-risk consumers, imposing incentive compatibility constraints to ensure that consumers select the right contracts, given their types. Given this, the insurer solves:

$\max_{(y1H,y2H),\ (y1L\ y2L)} \rho[(1-\pi_H)(Y_1- y_1^H) + \pi_H (Y_2 - y_2^H)] + (1-\rho) [(1-\pi_L)(Y_1- y_1^L) + \pi_L(Y_2 - y_2^L)]$
s.t.

$$(1-\pi_H)u(y_1^H) + \pi_H u(y_2^H) \geq (1-\pi_H)u(Y_1) + \pi_H u(Y_2) \quad \text{(participation constraint for H)}$$

$$(1-\pi_H)u(y_1^H) + \pi_H u(y_2^H) \geq (1-\pi_H)u(y_1^L) + \pi_H u(y_2^L) \quad \text{(incentive compatibility constraint for H)}$$

$$(1-\pi_L)u(y_1^L) + \pi_L u(y_2^L) \geq (1-\pi_L)u(Y_1) + \pi_L u(Y_2) \quad \text{(participation constraint for L)}$$

$$(1-\pi_L)u(y_1^L) + \pi_L u(y_2^L) \geq (1-\pi_L)u(y_1^H) + \pi_L u(y_2^H) \quad \text{(incentive compatibility constraint for L)}$$

Proposition 1 at Kreps 670: At the solution, the participation constraint for L and the incentive compatibility constraint for H are binding, and the high-risk contract (y_1^H,y_2^H) has full insurance, with $y_1^H=y_2^H$. Sketch of proof (see Kreps 670-674 for details): (i) The participation constraint for L binds, because you can't have both participation constraints slack, and the one for L binds before the one for H, "because" low-risk consumers need insurance less than high-risk consumers (Figure 18.5(b) at Kreps 671). (ii) The incentive compatibility constraint for H binds, "because" first-best contracts would make high-risk consumers want the low-risk contract. (iii) Thus, the solution of the principal's problem is as in Figure 18.6 at Kreps 674. Just where it lies on the 45˚ line depends on ρ: If ρ is near 1 (0), it will be near (or maybe at) the ideal contract for high-(low-)risk consumers.

Problems 14.B.1-8 at MWG 507-508 and 14.C.1-9 at MWG 508-510; and problems 1-2 and 4-7 at Kreps 616-623

9. Incentives and Mechanism Design

MWG 857-910; Kreps 661-703; McMillan 133-159

Principal-agent analysis is leading case of mechanism design, foundation for all that follows:

Analyses of public goods provision and public projects, quasilinear preferences and the pivot mechanism (MWG 861-862, 876-880; Kreps 704-712)

Allocation of indivisible goods (MWG 862-864)

Optimal auctions and revenue equivalence (MWG 865-866, 889-891)

Bilateral exchange and the Myerson-Satterthwaite Theorem (MWG 894-910, Kreps 680-703)

Dominant strategy and Bayesian implementation of social choice rules (MWG 857-897)

Problems at MWG 918-925; and problems 1-5 at Kreps 715-717