Draft chapter from *An introduction to game theory* by Martin J. Osborne. Version: 2002/7/23.
Martin.Osborne@utoronto.ca
http://www.economics.utoronto.ca/osborne
Copyright © 1995–2002 by Martin J. Osborne. All rights reserved. No part of this book may be reproduced by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from Oxford University Press, except that one copy of up to six chapters may be made by any individual for private study.

5 Extensive Games with Perfect Information: Theory

Prerequisite: Chapters 1 and 2.

THE MODEL of a strategic game suppresses the sequential structure of decision-making. When applying the model to situations in which decision-makers move sequentially, we assume that each decision-maker chooses her plan of action once and for all; she is committed to this plan, which she cannot modify as events unfold. The model of an extensive game, by contrast, describes the sequential structure of decision-making explicitly, allowing us to study situations in which each decision-maker is free to change her mind as events unfold.

In this chapter and the next two we study a model in which each decision-maker is always fully informed about all previous actions. In Chapter 10 we study a more general model, which allows each decision-maker, when taking an action, to be imperfectly informed about previous actions.

5.1 Extensive games with perfect information

5.1.1 Definition

To describe an extensive game with perfect information, we need to specify the set of players and their preferences, as for a strategic game (Definition 11.1). In addition, we need to specify the order of the players' moves and the actions each player may take at each point. We do so by specifying the set of all sequences of actions that can possibly occur, together with the player who moves at each point in each sequence. We refer to each possible sequence of actions as a *terminal history* and to the function that gives the player who moves at each point in each terminal history as the *player function*. That is, an extensive game has four components:

- players
- terminal histories
- player function
- preferences for the players.

Before giving precise definitions of these components, I give an example that illustrates them informally.

◆ EXAMPLE 152.1 (Entry game) An incumbent faces the possibility of entry by a challenger. (The challenger may, for example, be a firm considering entry into an industry currently occupied by a monopolist, a politician competing for the leadership of a party, or an animal considering competing for the right to mate with a congener of the opposite sex.) The challenger may enter or not. If it enters, the incumbent may either acquiesce or fight.

We may model this situation as an extensive game with perfect information in which the terminal histories are $(In, Acquiesce)$, $(In, Fight)$, and Out, and the player function assigns the challenger to the start of the game and the incumbent to the history In.

At the start of an extensive game, and after any sequence of events, a player chooses an action. The sets of actions available to the players are not, however, given explicitly in the description of the game. Instead, the description of the game specifies the set of terminal histories and the player function, from which we can deduce the available sets of actions.

In the entry game, for example, the actions available to the challenger at the start of the game are In and Out, because these actions (and no others) begin terminal histories, and the actions available to the incumbent are $Acquiesce$ and $Fight$, because these actions (and no others) follow In in terminal histories. More generally, suppose that (C, D) and (C, E) are terminal histories and the player function assigns player 1 to the start of the game and player 2 to the history C. Then two of the actions available to player 2 after player 1 chooses C at the start of the game are D and E.

The terminal histories of a game are specified as a set of sequences. But not every set of sequences is a legitimate set of terminal histories. If (C, D) is a terminal history, for example, there is no sense in specifying C as a terminal history: the fact that (C, D) is terminal implies that after C is chosen at the start of the game, some player may choose D, so that the action C does not end the game. More generally, a sequence that is a *proper subhistory* of a terminal history cannot itself be a terminal history. This restriction is the only one we need to impose on a set of sequences in order that the set be interpretable as a set of terminal histories.

To state the restriction precisely, define the **subhistories** of a finite sequence $(a^1, a^2, \ldots, a^k)$ of actions to be the empty sequence consisting of no actions, denoted $\varnothing$ (representing the start of the game), and all sequences of the form $(a^1, a^2, \ldots, a^m)$

where $1 \leq m \leq k$. (In particular, the entire sequence is a subhistory of itself.) Similarly, define the **subhistories** of an infinite sequence $(a^1, a^2, \ldots)$ of actions to be the empty sequence $\varnothing$, every sequence of the form $(a^1, a^2, \ldots, a^m)$ where m is a positive integer, and the entire sequence $(a^1, a^2, \ldots)$. A subhistory not equal to the entire sequence is called a **proper subhistory**. A sequence of actions that is a subhistory of some terminal history is called simply a **history**.

In the entry game in Example 152.1, the subhistories of $(In, Acquiesce)$ are the empty history $\varnothing$ and the sequences In and $(In, Acquiesce)$; the proper subhistories are the empty history and the sequence In.

▶ DEFINITION 153.1 (*Extensive game with perfect information*) An **extensive game with perfect information** consists of

- a set of **players**
- a set of sequences (**terminal histories**) with the property that no sequence is a proper subhistory of any other sequence
- a function (the **player function**) that assigns a player to every sequence that is a proper subhistory of some terminal history
- for each player, **preferences** over the set of terminal histories.

The set of terminal histories is the set of all sequences of actions that may occur; the player assigned by the player function to any history h is the player who takes an action after h.

As for a strategic game, we may specify a player's preferences by giving a payoff function that represents them (see Section 1.2.2). In some situations an outcome is associated with each terminal history, and the players' preferences are naturally defined over these outcomes, rather than directly over the terminal histories. For example, if we are modeling firms choosing prices then we may think in terms of each firm's caring about its profit—the outcome of a profile of prices—rather than directly about the profile of prices. However, any preferences over outcomes (e.g. profits) may be translated into preferences over terminal histories (e.g. sequences of prices). In the general definition, outcomes are conveniently identified with terminal histories and preferences are defined directly over these histories, avoiding the need for an additional element in the specification of the game.

◆ EXAMPLE 153.2 (Entry game) In the situation described in Example 152.1, suppose that the best outcome for the challenger is that it enters and the incumbent acquiesces, and the worst outcome is that it enters and the incumbent fights, whereas the best outcome for the incumbent is that the challenger stays out, and the worst outcome is that it enters and there is a fight. Then the situation may be modeled as the following extensive game with perfect information.

Players The challenger and the incumbent.

Terminal histories $(In, Acquiesce)$, $(In, Fight)$, and Out.

Player function $P(\varnothing) = Challenger$ and $P(In) = Incumbent$.

Preferences The challenger's preferences are represented by the payoff function u_1 for which $u_1(In, Acquiesce) = 2$, $u_1(Out) = 1$, and $u_1(In, Fight) = 0$, and the incumbent's preferences are represented by the payoff function u_2 for which $u_2(Out) = 2$, $u_2(In, Acquiesce) = 1$, and $u_2(In, Fight) = 0$.

This game is readily illustrated in a diagram. The small circle at the top of Figure 154.1 represents the empty history (the start of the game). The label above this circle indicates that the challenger chooses an action at the start of the game ($P(\varnothing) = Challenger$). The two branches labeled *In* and *Out* represent the challenger's choices. The segment labeled *In* leads to a small disk, where it is the incumbent's turn to choose an action ($P(In) = Incumbent$) and her choices are *Acquiesce* and *Fight*. The pair of numbers beneath each terminal history gives the players' payoffs to that history, with the challenger's payoff listed first. (The players' payoffs may be given in any order. For games like this one, in which the players move in a well-defined order, I generally list the payoffs in that order. For games in which the players' names are 1, 2, 3, and so on, I list the payoffs in the order of their names.)

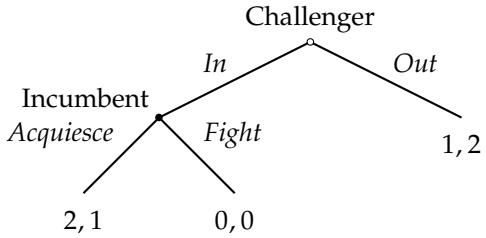

Figure 154.1 The entry game of Example 153.2. The challenger's payoff is the first number in each pair.

Definition 153.1 does not directly specify the sets of actions available to the players at their various moves. As I discussed briefly before the definition, we can deduce these sets from the set of terminal histories and the player function. If, for some nonterminal history h, the sequence (h, a) is a history, then a is one of the actions available to the player who moves after h. Thus the set of all actions available to the player who moves after h is

$$A(h) = \{a: (h, a) \text{ is a history}\}. \tag{154.1}$$

For example, for the game in Figure 154.1, the histories are $\varnothing$, In, Out, $(In, Acquiesce)$, and $(In, Fight)$. Thus the set of actions available to the player who moves at the start of the game, namely the challenger, is $A(\varnothing) = \{In, Out\}$, and the set of actions available to the player who moves after the history In, namely the incumbent, is $A(In) = \{Acquiesce, Fight\}$.

? EXERCISE 154.2 (Examples of extensive games with perfect information)

a. Represent in a diagram like Figure 154.1 the two-player extensive game with perfect information in which the terminal histories are (C, E), (C, F), (D, G),

and (D, H), the player function is given by $P(\varnothing) = 1$ and $P(C) = P(D) = 2$, player 1 prefers (C, F) to (D, G) to (C, E) to (D, H), and player 2 prefers (D, G) to (C, F) to (D, H) to (C, E).

b. Write down the set of players, set of terminal histories, player function, and players' preferences for the game in Figure 158.2.

c. The political figures Rosa and Ernesto have to choose a location for a party congress. The options are Berlin (B) or Havana (H). They choose sequentially. A third person, Karl, determines who chooses first. Both Rosa and Ernesto care only about the actions they choose, not about who chooses first. Rosa prefers the outcome in which both she and Ernesto choose B to that in which they both choose H, and prefers this outcome to either of the ones in which she and Ernesto choose different actions; she is indifferent between these last two outcomes. Ernesto's preferences differ from Rosa's in that the roles of B and H are reversed. Karl's preferences are the same as Ernesto's. Model this situation as an extensive game with perfect information. (Specify the components of the game and represent the game in a diagram.)

Definition 153.1 allows terminal histories to be infinitely long. Thus we can use the model of an extensive game to study situations in which the participants do not consider any particular fixed horizon when making decisions. If the length of the longest terminal history is in fact finite, we say that the game has a **finite horizon**.

Even a game with a finite horizon may have infinitely many terminal histories, because some player has infinitely many actions after some history. If a game has a finite horizon *and* finitely many terminal histories we say it is **finite**. Note that a game that is not finite cannot be represented in a diagram like Figure 154.1, because such a figure allows for only finitely many branches.

An extensive game with perfect information models a situation in which each player, when choosing an action, knows all actions chosen previously (has *perfect information*), and always moves alone (rather than simultaneously with other players). Some economic and political situations that the model encompasses are discussed in the next chapter. The competition between interest groups courting legislators is one example. This situation may be modeled as an extensive game in which the groups sequentially offer payments to induce the legislators to vote for their favorite version of a bill (Section 6.3). A race (between firms developing a new technology, or between directors making competing movies, for instance), is another example. This situation is modeled as an extensive game in which the parties alternately decide how much effort to expend (Section 6.4). Parlor games such as chess, ticktacktoe, and go, in which there are no random events, the players move sequentially, and each player always knows all actions taken previously, may also be modeled as extensive games with perfect information (see the box on page 176).

In Section 7.1 I discuss a more general notion of an extensive game in which players may move simultaneously, though each player, when choosing an action,

still knows all previous actions. In Chapter 10 I discuss a much more general notion that allows arbitrary patterns of information. In each case I sometimes refer to the object under consideration simply as an "extensive game".

5.1.2 Solutions

In the entry game in Figure 154.1, it seems clear that the challenger will enter and the incumbent will subsequently acquiesce. The challenger can reason that if it enters then the incumbent will acquiesce, because doing so is better for the incumbent than fighting. Given that the incumbent will respond to entry in this way, the challenger is better off entering.

This line of argument is called *backward induction*. Whenever a player has to move, she deduces, for each of her possible actions, the actions that the players (including herself) will subsequently rationally take, and chooses the action that yields the terminal history she most prefers.

While backward induction may be applied to the game in Figure 154.1, it cannot be applied to every extensive game with perfect information. Consider, for example, the variant of this game shown in Figure 156.1, in which the incumbent's payoff to the terminal history $(In, Fight)$ is 1 rather than 0. If, in the modified game, the challenger enters, the incumbent is indifferent between acquiescing and fighting. Backward induction does not tell the challenger what the incumbent will do in this case, and thus leaves open the question of which action the challenger should choose. Games with infinitely long histories present another difficulty for backward induction: they have no end from which to start the induction. The generalization of an extensive game with perfect information that allows for simultaneous moves (studied in Chapter 7) poses yet another problem: when players move simultaneously we cannot in general straightforwardly deduce each player's optimal action. (As in a strategic game, each player's best action depends on the other players' actions.)

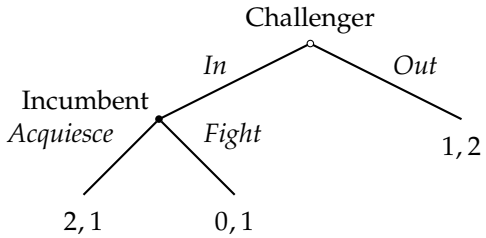

Figure 156.1 A variant of the entry game of Figure 154.1. The challenger's payoff is the first number in each pair.

Another approach to defining equilibrium takes off from the notion of Nash equilibrium. It seeks to model patterns of behavior that can persist in a steady state. The resulting notion of equilibrium applies to all extensive games with perfect information. Because the idea of backward induction is more limited, and the principles behind the notion of Nash equilibrium have been established in

previous chapters, I begin by discussing the steady state approach. In games in which backward induction is well-defined, this approach turns out to lead to the backward induction outcome, so that there is no conflict between the two ideas.

5.2 Strategies and outcomes

5.2.1 Strategies

A key concept in the study of extensive games is that of a *strategy*. A player's strategy specifies the action the player chooses for *every* history after which it is her turn to move.

▶ DEFINITION 157.1 (*Strategy*) A **strategy** of player i in an extensive game with perfect information is a function that assigns to each history h after which it is player i's turn to move (i.e. $P(h) = i$, where P is the player function) an action in $A(h)$ (the set of actions available after h).

Consider the game in Figure 157.1.

- Player 1 moves only at the start of the game (i.e. after the empty history), when the actions available to her are C and D. Thus she has two strategies: one that assigns C to the empty history, and one that assigns D to the empty history.
- Player 2 moves after both the history C and the history D. After the history C the actions available to her are E and F, and after the history D the actions available to her are G and H. Thus a strategy of player 2 is a function that assigns either E or F to the history C, and either G or H to the history D. That is, player 2 has *four* strategies, which are shown in Figure 158.1.

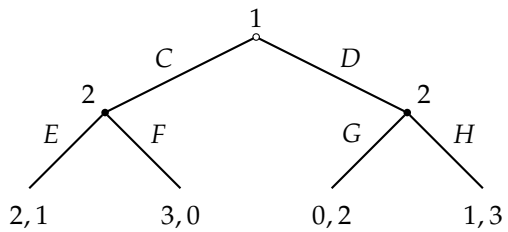

Figure 157.1 An extensive game with perfect information.

I refer to the strategies of player 1 in this game simply as C and D, and to the strategies of player 2 simply as EG, EH, FG, and FH. For many other finite games I use a similar shorthand: I write a player's strategy as a list of actions, one for each history after which it is the player's turn to move. In general I write the actions in the order in which they occur in the game, and, if they are available at the same "stage", from left to right as they appear in the diagram of the game. When the

	Action assigned to history C	Action assigned to history D
Strategy 1	E	G
Strategy 2	E	H
Strategy 3	F	G
Strategy 4	F	H

Figure 158.1 The four strategies of player 2 in the game in Figure 157.1.

meaning of a list of actions is unclear, I explicitly give the history after which each action is taken.

Each of player 2's strategies in the game in Figure 157.1 may be interpreted as a plan of action or contingency plan: it specifies what player 2 does *if* player 1 chooses C, *and* what she does *if* player 1 chooses D. In every game, a player's strategy provides sufficient information to determine her *plan of action*: the actions she intends to take, *whatever* the other players do. In particular, if a player appoints an agent to play the game for her, and tells the agent her strategy, then the agent has enough information to carry out her wishes, *whatever* actions the other players take.

In some games some players' strategies are *more* than plans of action. Consider the game in Figure 158.2. Player 1 moves both at the start of the game and after the history (C, E). In each case she has two actions, so she has *four* strategies: CG (i.e. choose C at the start of the game and G after the history (C, E)), CH, DG, and DH. In particular, each strategy specifies an action after the history (C, E) *even if it specifies the action D at the beginning of the game*, in which case the history (C, E) does not occur! The point is that Definition 157.1 requires that a strategy of any player i specify an action for *every* history after which it is player i's turn to move, *even for histories that, if the strategy is followed, do not occur*.

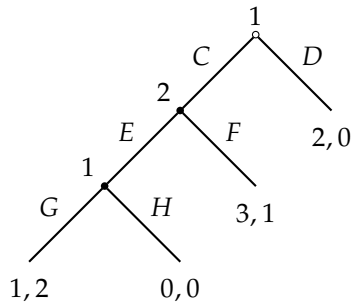

Figure 158.2 An extensive game in which player 1 moves both before and after player 2.

In view of this point and the fact that "strategy" is a synonym for "plan of action" in everyday language, you may regard the word "strategy" as inappropriate for the concept in Definition 157.1. You are right. You may also wonder why we cannot restrict attention to plans of action.

For the purposes of the notion of Nash equilibrium (discussed in the next section), we *could* in fact work with plans of action rather than strategies. But, as we shall see, the notion of Nash equilibrium for an extensive game is not satisfactory; the concept we adopt depends on the players' full strategies. When discussing this concept (in Section 5.4.4) I elaborate on the interpretation of a strategy. At the moment, you may think of a player's strategy as a plan of what to do, whatever the other players do, both if the player carries out her intended actions, and also if she makes mistakes. For example, we can interpret the strategy DG of player 1 in the game in Figure 158.2 to mean "I intend to choose D, but if I make a mistake and choose C instead then I will subsequently choose G". (Because the notion of Nash equilibrium depends only on plans of action, I could delay the definition of a strategy to the start of Section 5.4. I do not do so because the notion of a strategy is central to the study of extensive games, and its precise definition is much simpler than that of a plan of action.)

? EXERCISE 159.1 (Strategies in extensive games) What are the strategies of the players in the entry game (Example 153.2)? What are Rosa's strategies in the game in Exercise 154.2c?

5.2.2 Outcomes

A strategy profile determines the terminal history that occurs. Denote the strategy profile by s and the player function by P. At the start of the game player $P(\varnothing)$ moves. Her strategy is $s_{P(\varnothing)}$, and she chooses the action $s_{P(\varnothing)}(\varnothing)$. Denote this action by a^1. If the history a^1 is not terminal, player $P(a^1)$ moves next. Her strategy is $s_{P(a^1)}$, and she chooses the action $s_{P(a^1)}(a^1)$. Denote this action by a^2. If the history (a^1, a^2) is not terminal, then again the player function specifies whose turn it is to move, and that player's strategy specifies the action she chooses. The process continues until a terminal history is constructed. We refer to this terminal history as the **outcome of** s, and denote it $O(s)$.

In the game in Figure 158.2, for example, the outcome of the strategy pair (DG, E) is the terminal history D, and the outcome of (CH, E) is the terminal history (C, E, H).

Note that the outcome $O(s)$ of the strategy profile s depends only on the players' plans of action, not their full strategies. That is, to determine $O(s)$ we do *not* need to refer to any component of any player's strategy that specifies her actions after histories precluded by that strategy.

5.3 Nash equilibrium

As for strategic games, we are interested in notions of equilibrium that model the players' behavior in a steady state. That is, we look for patterns of behavior with the property that if every player knows every other player's behavior, she has no reason to change her own behavior. I start by defining a Nash equilibrium: a strat-

egy profile from which no player wishes to deviate, given the other players' strategies. The definition is an adaptation of that of a Nash equilibrium in a strategic game (21.1).

▶ DEFINITION 160.1 (*Nash equilibrium of extensive game with perfect information*) The strategy profile s^* in an extensive game with perfect information is a **Nash equilibrium** if, for every player i and every strategy r_i of player i, the terminal history $O(s^*)$ generated by s^* is at least as good according to player i's preferences as the terminal history $O(r_i, s^*_{-i})$ generated by the strategy profile (r_i, s^*_{-i}) in which player i chooses r_i while every other player j chooses s^*_j. Equivalently, for each player i,

$$u_i(O(s^*)) \geq u_i(O(r_i, s^*_{-i})) \text{ for every strategy } r_i \text{ of player } i,$$

where u_i is a payoff function that represents player i's preferences and O is the outcome function of the game.

One way to find the Nash equilibria of an extensive game in which each player has finitely many strategies is to list each player's strategies, find the outcome of each strategy profile, and analyze this information as for a strategic game. That is, we construct the following strategic game, known as the **strategic form** of the extensive game.

Players The set of players in the extensive game.

Actions Each player's set of actions is her set of strategies in the extensive game.

Preferences Each player's payoff to each action profile is her payoff to the terminal history generated by that action profile in the extensive game.

From Definition 160.1 we see that

> the set of Nash equilibria of any extensive game with perfect information is the set of Nash equilibria of its strategic form.

◆ EXAMPLE 160.2 (Nash equilibria of the entry game) In the entry game in Figure 154.1, the challenger has two strategies, *In* and *Out*, and the incumbent has two strategies, *Acquiesce* and *Fight*. The strategic form of the game is shown in Figure 161.1. We see that it has two Nash equilibria: $(In, Acquiesce)$ and $(Out, Fight)$. The first equilibrium is the pattern of behavior isolated by backward induction, discussed at the start of Section 5.1.2.

In the second equilibrium the challenger always chooses *Out*. This strategy is optimal given the incumbent's strategy to fight in the event of entry. Further, the incumbent's strategy *Fight* is optimal given the challenger's strategy: the challenger chooses *Out*, so whether the incumbent plans to choose *Acquiesce* or *Fight* makes no difference to its payoff. Thus neither player can increase its payoff by choosing a different strategy, *given the other player's strategy*.

		Incumbent	
		Acquiesce	*Fight*
Challenger	*In*	2, 1	0, 0
	Out	1, 2	1, 2

Figure 161.1 The strategic form of the entry game in Figure 154.1.

Thinking about the extensive game in this example raises a question about the Nash equilibrium $(Out, Fight)$ that does not arise when thinking about the strategic form: how does the challenger know that the incumbent will choose *Fight* if it enters? We interpret the strategic game to model a situation in which, whenever the challenger plays the game, it observes the incumbent's action, even if it chooses *Out*. By contrast, we interpret the extensive game to model a situation in which a challenger that always chooses *Out* never observes the incumbent's action, because the incumbent never moves. In a strategic game, the rationale for the Nash equilibrium condition that each player's strategy be optimal given the other players' strategies is that in a steady state, each player's experience playing the game leads her belief about the other players' actions to be correct. This rationale does not apply to the Nash equilibrium $(Out, Fight)$ of the (extensive) entry game, because a challenger who always chooses *Out* never observes the incumbent's action after the history *In*.

We can escape from this difficulty in interpreting a Nash equilibrium of an extensive game by considering a slightly perturbed steady state in which, on rare occasions, nonequilibrium actions are taken (perhaps players make mistakes, or deliberately experiment), and the perturbations allow each player eventually to observe every other player's action after *every* history. Given such perturbations, each player eventually learns the other players' entire strategies.

Interpreting the Nash equilibrium $(Out, Fight)$ as such a perturbed steady state, however, we run into another problem. On those (rare) occasions when the challenger enters, the subsequent behavior of the incumbent to fight is not a steady state in the remainder of the game: if the challenger enters, the incumbent is better off acquiescing than fighting. That is, the Nash equilibrium $(Out, Fight)$ does not correspond to a *robust* steady state of the extensive game.

Note that the extensive game embodies the assumption that the incumbent cannot commit, at the beginning of the game, to fight if the challenger enters; it is free to choose either *Acquiesce* or *Fight* in this event. If the incumbent *could* commit to fight in the event of entry then the analysis would be different. Such a commitment would induce the challenger to stay out, an outcome that the incumbent prefers. In the absence of the possibility of the incumbent's making a commitment, we might think of its *announcing* at the start of the game that it intends to fight; but such a *threat* is not credible, because after the challenger enters the incumbent's only incentive is to acquiesce.

? EXERCISE 162.1 (Nash equilibria of extensive games) Find the Nash equilibria of the games in Exercise 154.2*a* and Figure 158.2. (When constructing the strategic form of each game, be sure to include *all* the strategies of each player.)

? EXERCISE 162.2 (Voting by alternating veto) Two people select a policy that affects them both by alternately vetoing policies until only one remains. First person 1 vetoes a policy. If more than one policy remains, person 2 then vetoes a policy. If more than one policy still remains, person 1 then vetoes another policy. The process continues until a single policy remains unvetoed. Suppose there are three possible policies, X, Y, and Z, person 1 prefers X to Y to Z, and person 2 prefers Z to Y to X. Model this situation as an extensive game and find its Nash equilibria.

5.4 Subgame perfect equilibrium

5.4.1 Definition

The notion of Nash equilibrium ignores the sequential structure of an extensive game; it treats strategies as choices made once and for all before play begins. Consequently, as we saw in the previous section, the steady state to which a Nash equilibrium corresponds may not be robust.

I now define a notion of equilibrium that models a robust steady state. This notion requires each player's strategy to be optimal, given the other players' strategies, not only at the start of the game, but after every possible history.

To define this concept, I first define the notion of a subgame. For any nonterminal history h, the *subgame* following h is the part of the game that remains after h has occurred. For example, the subgame following the history *In* in the entry game (Example 152.1) is the game in which the incumbent is the only player, and there are two terminal histories, *Acquiesce* and *Fight*.

▶ DEFINITION 162.3 (*Subgame of extensive game with perfect information*) Let Γ be an extensive game with perfect information, with player function P. For any nonterminal history h of Γ, the **subgame** $\Gamma(h)$ **following the history** h is the following extensive game.

Players The players in Γ.

Terminal histories The set of all sequences h' of actions such that (h, h') is a terminal history of Γ.

Player function The player $P(h, h')$ is assigned to each proper subhistory h' of a terminal history.

Preferences Each player prefers h' to h'' if and only if she prefers (h, h') to (h, h'') in Γ.

Note that the subgame following the initial history $\varnothing$ is the entire game. Every other subgame is called a *proper subgame*. Because there is a subgame for every nonterminal history, the number of subgames is equal to the number of nonterminal histories.

As an example, the game in Figure 157.1 has three nonterminal histories (the initial history, C, and D), and hence three subgames: the whole game (the part of the game following the initial history), the game following the history C, and the game following the history D. The two proper subgames are shown in Figure 163.1.

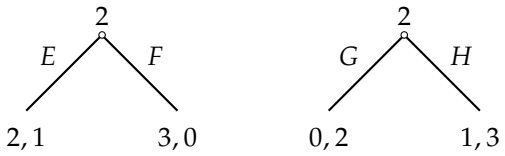

Figure 163.1 The two proper subgames of the extensive game in Figure 157.1.

The game in Figure 158.2 also has three nonterminal histories, and hence three subgames: the whole game, the game following the history C, and the game following the history (C, E). The two proper subgames are shown in Figure 163.2.

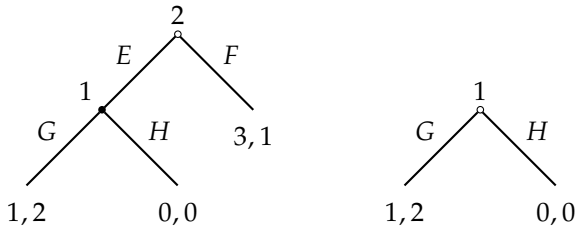

Figure 163.2 The two proper subgames of the extensive game in Figure 158.2.

? EXERCISE 163.1 (Subgames) Find all the subgames of the game in Exercise 154.2*c*.

In an equilibrium that corresponds to a perturbed steady state in which *every* history sometimes occurs, the players' behavior must correspond to a steady state in *every subgame*, not only in the whole game. Interpreting the actions specified by a player's strategy in a subgame to give the player's behavior if, possibly after a series of mistakes, that subgame is reached, this condition is embodied in the following informal definition.

> A *subgame perfect equilibrium* is a strategy profile s^* with the property that in no subgame can any player i do better by choosing a strategy different from s_i^*, given that every other player j adheres to s_j^*.

(Compare this definition with that of a Nash equilibrium of a strategic game, on page 20.)

For example, the Nash equilibrium $(Out, Fight)$ of the entry game (Example 152.1) is not a subgame perfect equilibrium because in the subgame following the history *In*, the strategy *Fight* is not optimal for the incumbent: *in this subgame*, the incumbent is better off choosing *Acquiesce* than it is choosing *Fight*. The Nash equilibrium

$(In, Acquiesce)$ *is* a subgame perfect equilibrium: each player's strategy is optimal, given the other player's strategy, both in the whole game, and in the subgame following the history In.

To define the notion of subgame perfect equilibrium precisely, we need a new piece of notation. Let h be a history and s a strategy profile. Suppose that h occurs (even though it is not necessarily consistent with s), and *afterwards* the players adhere to the strategy profile s. Denote the resulting terminal history by $O_h(s)$. That is, $O_h(s)$ is the terminal history consisting of h followed by the outcome generated in the subgame following h by the strategy profile induced by s in the subgame. Note that for any strategy profile s, we have $O_\varnothing(s) = O(s)$ (where $\varnothing$, as always, denotes the initial history).

As an example, consider again the entry game. Let s be the strategy profile $(Out, Fight)$ and let h be the history In. If h occurs, and *afterwards* the players adhere to s, the resulting terminal history is $O_h(s) = (In, Fight)$.

▶ DEFINITION 164.1 (*Subgame perfect equilibrium of extensive game with perfect information*) The strategy profile s^* in an extensive game with perfect information is a **subgame perfect equilibrium** if, for every player i, every history h after which it is player i's turn to move (i.e. $P(h) = i$), and every strategy r_i of player i, the terminal history $O_h(s^*)$ generated by s^* after the history h is at least as good according to player i's preferences as the terminal history $O_h(r_i, s^*_{-i})$ generated by the strategy profile (r_i, s^*_{-i}) in which player i chooses r_i while every other player j chooses s^*_j. Equivalently, for every player i and every history h after which it is player i's turn to move,

$$u_i(O_h(s^*)) \geq u_i(O_h(r_i, s^*_{-i})) \text{ for every strategy } r_i \text{ of player } i,$$

where u_i is a payoff function that represents player i's preferences and $O_h(s)$ is the terminal history consisting of h followed by the sequence of actions generated by s after h.

The important point in this definition is that each player's strategy is required to be optimal for *every* history after which it is the player's turn to move, not only at the start of the game as in the definition of a Nash equilibrium (160.1).

5.4.2 Subgame perfect equilibrium and Nash equilibrium

In a subgame perfect equilibrium every player's strategy is optimal, in particular, after the initial history (put $h = \varnothing$ in the definition, and remember that $O_\varnothing(s) = O(s)$). Thus

every subgame perfect equilibrium is a Nash equilibrium.

In fact, a subgame perfect equilibrium generates a Nash equilibrium in every subgame: if s^* is a subgame perfect equilibrium then, for any history h and player i, the strategy induced by s^*_i in the subgame following h is optimal given the strategies induced by s^*_{-i} in the subgame. Further, any strategy profile that generates a

Nash equilibrium in every subgame is a subgame perfect equilibrium, so that we can give the following alternative definition.

> A *subgame perfect equilibrium* is a strategy profile that induces a Nash equilibrium in every subgame.

In a Nash equilibrium every player's strategy is optimal, given the other players' strategies, in the whole game. As we have seen, it may *not* be optimal in some subgames. I claim, however, that it *is* optimal in any subgame that is reached when the players follow their strategies. Given this claim, the significance of the requirement in the definition of a subgame perfect equilibrium that each player's strategy be optimal after every history, relative to the requirement in the definition of a Nash equilibrium, is that each player's strategy be optimal after histories that do not occur if the players follow their strategies (like the history *In* when the challenger's action is *Out* at the beginning of the entry game).

To show my claim, suppose that s^* is a Nash equilibrium of a game in which you are player i. Then your strategy s_i^* is optimal given the other players' strategies s_{-i}^*. When the other players follow their strategies, there comes a point (possibly the start of the game) when you have to move for the first time. Suppose that at this point you follow your strategy s_i^*; denote the action you choose by C. Now, after having chosen C, should you change your strategy in the rest of the game, given that the other players will continue to adhere to their strategies? No! If you could do better by changing your strategy after choosing C—say by switching to the strategy s_i' in the subgame—then you could have done better at the start of the game by choosing the strategy that chooses C and then follows s_i'. That is, if your plan is optimal, given the other players' strategies, at the start of the game, and you stick to it, then you never want to change your mind after play begins, as long as the other players stick to their strategies. (The general principle is known as the *Principle of Optimality* in dynamic programming.)

5.4.3 Examples

◆ EXAMPLE 165.1 (Entry game) Consider again the entry game of Example 152.1, which has two Nash equilibria, $(In, Acquiesce)$ and $(Out, Fight)$. The fact that the Nash equilibrium $(Out, Fight)$ is not a subgame perfect equilibrium follows from the formal definition as follows. For $s^* = (Out, Fight)$, $i =$ Incumbent, $r_i = Acquiesce$, and $h = In$, we have $O_h(s^*) = (In, Fight)$ and $O_h(r_i, s_{-i}^*) = (In, Acquiesce)$, so that the inequality in the definition is violated: $u_i(O_h(s^*)) = 0$ and $u_i(O_h(r_i, s_{-i}^*)) = 1$.

The Nash equilibrium $(In, Acquiesce)$ *is* a subgame perfect equilibrium because (*a*) it is a Nash equilibrium, so that at the start of the game the challenger's strategy *In* is optimal, given the incumbent's strategy *Acquiesce*, and (*b*) after the history *In*, the incumbent's strategy *Acquiesce* in the subgame is optimal. In the language of the formal definition, let $s^* = (In, Acquiesce)$.

- The challenger moves after one history, namely $h = \varnothing$. We have $O_h(s^*) =$

$(In, Acquiesce)$ and hence for $i =$ challenger we have $u_i(O_h(s^*)) = 2$, whereas for the only other strategy of the challenger, $r_i = Out$, we have $u_i(O_h(r_i, s^*_{-i})) = 1$.

- The incumbent moves after one history, namely $h = In$. We have $O_h(s^*) = (In, Acquiesce)$ and hence for $i =$ incumbent we have $u_i(O_h(s^*)) = 1$, whereas for the only other strategy of the incumbent, $r_i = Fight$, we have $u_i(O_h(r_i, s^*_{-i})) = 0$.

Every subgame perfect equilibrium is a Nash equilibrium, so we conclude that the game has a unique subgame perfect equilibrium, $(In, Acquiesce)$.

◆ EXAMPLE 166.1 (Variant of entry game) Consider the variant of the entry game in which the incumbent is indifferent between fighting and acquiescing if the challenger enters (see Figure 156.1). This game, like the original game, has two Nash equilibria, $(In, Acquiesce)$ and $(Out, Fight)$. But now *both* of these equilibria are subgame perfect equilibria, because after the history *In* both *Fight* and *Acquiesce* are optimal for the incumbent.

In particular, the game has a steady state in which every challenger always chooses *In* and every incumbent always chooses *Acquiesce*. If you, as the challenger, were playing the game for the first time, you would probably regard the action *In* as "risky", because after the history *In* the incumbent is indifferent between *Acquiesce* and *Fight*, and you prefer the terminal history *Out* to the terminal history $(In, Fight)$. Indeed, as discussed in Section 5.1.2, backward induction does not yield a clear solution of this game. But the subgame perfect equilibrium $(In, Acquiesce)$ corresponds to a perfectly reasonable steady state. If you had played the game hundreds of times against opponents drawn from the same population, and on every occasion your opponent had chosen *Acquiesce*, you could reasonably expect your next opponent to choose *Acquiesce*, and thus optimally choose *In*.

? EXERCISE 166.2 (Checking for subgame perfect equilibria) Which of the Nash equilibria of the game in Figure 158.2 are subgame perfect?

5.4.4 Interpretation

A Nash equilibrium of a strategic game corresponds to a steady state in an idealized setting in which the participants in each play of the game are drawn randomly from a collection of populations (see Section 2.6). The idea is that each player's long experience playing the game leads her to correct beliefs about the other players' actions; given these beliefs her equilibrium action is optimal.

A subgame perfect equilibrium of an extensive game corresponds to a slightly perturbed steady state, in which all players, on rare occasions, take nonequilibrium actions, so that after long experience each player forms correct beliefs about the other players' entire strategies, and thus knows how the other players will behave in every subgame. Given these beliefs, no player wishes to deviate from her strategy either at the start of the game or after *any* history.

This interpretation of a subgame perfect equilibrium, like the interpretation of a Nash equilibrium as a steady state, does not require a player to know the other players' preferences, or to think about the other players' rationality. It entails interpreting a strategy as a plan specifying a player's actions not only after histories consistent with the strategy, but also after histories that result when the player chooses arbitrary alternative actions, perhaps because she makes mistakes or deliberately experiments.

The subgame perfect equilibria of some extensive game can be given other interpretations. In some cases, one alternative interpretation is particularly attractive. Consider an extensive game with perfect information in which each player has a unique best action at every history after which it is her turn to move, and the horizon is finite. In such a game, a player who knows the other players' preferences and knows that the other players are rational may use backward induction to deduce her optimal strategy, as discussed in Section 5.1.2. Thus we can interpret a subgame perfect equilibrium as the outcome of the players' rational calculations about each other's strategies.

This interpretation of a subgame perfect equilibrium entails an interpretation of a strategy different from the one that fits the steady state interpretation. Consider, for example, the game in Figure 158.2. When analyzing this game, player 1 must consider the consequences of choosing C. Thus she must think about player 2's action after the history C, and hence must form a belief about what player 2 thinks she (player 1) will do after the history (C, E). The component of her strategy that specifies her action after this history reflects this belief. For instance, the strategy DG means that player 1 chooses D at the start of the game and believes that were she to choose C, player 2 would believe that after the history (C, E) she would choose G. In an arbitrary game, the interpretation of a subgame perfect equilibrium as the outcome of the players' rational calculations about each other's strategies entails interpreting the components of a player's strategy that assign actions to histories inconsistent with other parts of the strategy as specifying the player's belief about the other players' beliefs about what the player will do if one of these histories occurs.

This interpretation of a subgame perfect equilibrium is not free of difficulties, which are discussed in Section 7.7. Further, the interpretation is not tenable in games in which some player has more than one optimal action after some history, or in the more general extensive games considered in Section 7.1 and Chapter 10. Nevertheless, in some of the games studied in this chapter and the next it is an appealing alternative to the steady state interpretation. Further, an extension of the procedure of backward induction can be used to find all subgame perfect equilibria of finite horizon games, as we shall see in the next section. (This extension cannot be given an appealing behavioral interpretation in games in which some player has more than one optimal action after some history.)

5.5 Finding subgame perfect equilibria of finite horizon games: backward induction

We found the subgame perfect equilibria of the games in Examples 165.1 and 166.1 by finding the Nash equilibria of the games and checking whether each of these equilibria is subgame perfect. In a game with a finite horizon the set of subgame perfect equilibria may be found more directly by using an extension of the procedure of backward induction discussed briefly in Section 5.1.2.

Define the *length of a subgame* to be the length of the longest history in the subgame. (The lengths of the subgames in Figure 163.2, for example, are 2 and 1.) The procedure of backward induction works as follows. We start by finding the optimal actions of the players who move in the subgames of length 1 (the "last" subgames). Then, taking these actions as given, we find the optimal actions of the players who move first in the subgames of length 2. We continue working back to the beginning of the game, at each stage k finding the optimal actions of the players who move at the start of the subgames of length k, given the optimal actions we have found in all shorter subgames.

At each stage k of this procedure, the optimal actions of the players who move at the start of the subgames of length k are easy to determine: they are simply the actions that yield the players the highest payoffs, given the optimal actions in all shorter subgames.

Consider, for example, the game in Figure 169.1.

- First consider subgames of length 1. The game has two such subgames, in both of which player 2 moves. In the subgame following the history C, player 2's optimal action is E, and in the subgame following the history D, her optimal action is H.
- Now consider subgames of length 2. The game has one such subgame, namely the entire game, at the start of which player 1 moves. Given the optimal actions in the subgames of length 1, player 1's choosing C at the start of the game yields her a payoff of 2, whereas her choosing D yields her a payoff of 1. Thus player 1's optimal action at the start of the game is C.

The game has no subgame of length greater than 2, so the procedure of backward induction yields the strategy pair (C, EH).

As another example, consider the game in Figure 158.2. We first deduce that in the subgame of length 1 following the history (C, E), player 1 chooses G; then that at the start of the subgame of length 2 following the history C, player 2 chooses E; then that at the start of the whole game, player 1 chooses D. Thus the procedure of backward induction in this game yields the strategy pair (DG, E).

In any game in which this procedure selects a single action for the player who moves at the start of each subgame, the strategy profile thus selected is the unique subgame perfect equilibrium of the game. (You should find this result very plausible, though a complete proof is not trivial.)

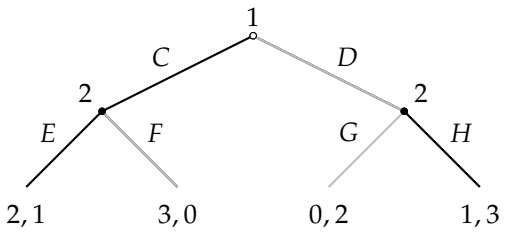

Figure 169.1 A game illustrating the procedure of backward induction. The actions selected by backward induction are indicated in black.

What happens in a game in which at the start of some subgames more than one action is optimal? In such a game an extension of the procedure of backward induction locates all subgame perfect equilibrium. This extension traces back *separately* the implications for behavior in the longer subgames of *every combination* of optimal actions in the shorter subgames.

Consider, for example, the game in Figure 170.1.

- The game has three subgames of length one, in each of which player 2 moves. In the subgames following the histories C and D, player 2 is indifferent between her two actions. In the subgame following the history E, player 2's unique optimal action is K. Thus there are *four* combinations of player 2's optimal actions in the subgames of length 1: FHK, FIK, GHK, and GIK (where the first component in each case is player 2's action after the history C, the second component is her action after the history D, and the third component is her action after the history E).

- The game has a single subgame of length two, namely the whole game, in which player 1 moves first. We now consider player 1's optimal action in this game for *every combination* of the optimal actions of player 2 in the subgames of length 1.

  - For the combinations FHK and FIK of optimal actions of player 2, player 1's optimal action at the start of the game is C.
  - For the combination GHK of optimal actions of player 2, the actions C, D, and E are all optimal for player 1.
  - For the combination GIK of optimal actions of player 2, player 1's optimal action at the start of the game is D.

Thus the strategy pairs isolated by the procedure are (C, FHK), (C, FIK), (C, GHK), (D, GHK), (E, GHK), and (D, GIK).

The procedure, which for simplicity I refer to simply as **backward induction**, may be described compactly for an arbitrary game as follows.

- Find, for each subgame of length 1, the set of optimal actions of the player who moves first. Index the subgames by j, and denote by $S_j^*(1)$ the set of

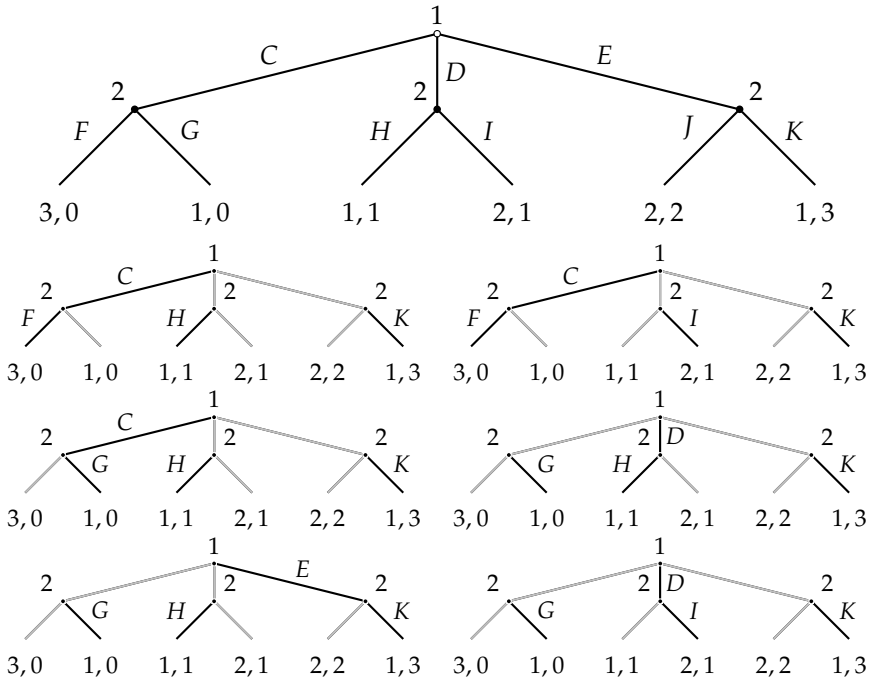

Figure 170.1 A game in which the first-mover in some subgames has multiple optimal actions. The top diagram shows the full game. The six small diagrams illustrate the six subgame perfect equilibria; in each case, the actions specified by the equilibrium strategies are indicated by black lines, and the remaining actions are indicated by gray lines.

optimal actions in subgame j. (If the player who moves first in subgame j has a unique optimal action, then $S_j^*(1)$ contains a single action.)

- For each combination of actions consisting of one from each set $S_j^*(1)$, find, for each subgame of length two, the set of optimal actions of the player who moves first. The result is a set of strategy profiles for each subgame of length two. Denote by $S_\ell^*(2)$ the set of strategy profiles in subgame ℓ.

- Continue by examining successively longer subgames until reaching the start of the game. At each stage k, for each combination of strategy profiles consisting of one from each set $S_p^*(k-1)$ constructed in the previous stage, find, for each subgame of length k, the set of optimal actions of the player who moves first, and hence a set of strategy profiles for each subgame of length k.

The set of strategy profiles that this procedure yields for the whole game is the set of subgame perfect equilibria of the game.

■ PROPOSITION 170.1 (Subgame perfect equilibrium of finite horizon games and backward induction) *The set of subgame perfect equilibria of a finite horizon extensive game with perfect information is equal to the set of strategy profiles isolated by the procedure of backward induction.*

You should find this result, like my claim for games in which the player who moves at the start of every subgame has a single optimal action, very plausible, though again a complete proof is not trivial.

In the terminology of my description of the general procedure, the analysis for the game in Figure 170.1 is as follows. Number the subgames of length one from left to right. Then we have $S_1^*(1) = \{F, G\}$, $S_2^*(1) = \{H, I\}$, and $S_3^*(1) = \{K\}$. There are four lists of actions consisting of one action from each set: *FHK*, *FIK*, *GHK*, and *GIK*. For *FHK* and *FIK*, the action *C* of player 1 is optimal at the start of the game; for *GHK* the actions *C*, *D*, and *E* are all optimal; and for *GIK* the action *D* is optimal. Thus the set $S^*(2)$ of strategy profiles consists of (C, FHK), (C, FIK), (C, GHK), (D, GHK), (E, GHK), and (D, GIK). There are no longer subgames, so this set of strategy profiles is the set of subgame perfect equilibria of the game.

Each example I have presented so far in this section is a finite game—that is, a game that not only has a finite horizon, but also a finite number of terminal histories. In such a game, the player who moves first in any subgame has finitely many actions; at least one action is optimal. Thus in such a game the procedure of backward induction isolates at least one strategy profile. Using Proposition 170.1, we conclude that every finite game has a subgame perfect equilibrium.

■ PROPOSITION 171.1 (Existence of subgame perfect equilibrium) *Every finite extensive game with perfect information has a subgame perfect equilibrium.*

Note that this result does *not* claim that a finite extensive game has a *single* subgame perfect equilibrium. (As we have seen, the game in Figure 170.1, for example, has more than one subgame perfect equilibrium.)

A finite horizon game in which some player does not have finitely many actions after some history may or may not possess a subgame perfect equilibrium. A simple example of a game that does not have a subgame perfect equilibrium is the trivial game in which a single player chooses a number *less than* 1 and receives a payoff equal to the number she chooses. There is no greatest number less than one, so the single player has no optimal action, and thus the game has no subgame perfect equilibrium.

? EXERCISE 171.2 (Finding subgame perfect equilibria) Find the subgame perfect equilibria of the games in parts *a* and *c* of Exercise 154.2, and in Figure 171.1.

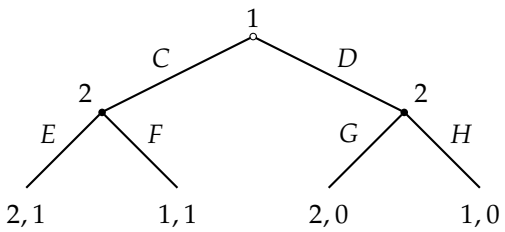

Figure 171.1 One of the games for Exercise 171.2.

? EXERCISE 172.1 (Voting by alternating veto) Find the subgame perfect equilibria of the game in Exercise 162.2. Does the game have any Nash equilibrium that is not a subgame perfect equilibrium? Is any outcome generated by a Nash equilibrium not generated by any subgame perfect equilibrium? Consider variants of the game in which player 2's preferences may differ from those specified in Exercise 162.2. Are there any preferences for which the outcome in a subgame perfect equilibrium of the game in which player 1 moves first differs from the outcome in a subgame perfect equilibrium of the game in which player 2 moves first?

? EXERCISE 172.2 (Burning a bridge) Army 1, of country 1, must decide whether to attack army 2, of country 2, which is occupying an island between the two countries. In the event of an attack, army 2 may fight, or retreat over a bridge to its mainland. Each army prefers to occupy the island than not to occupy it; a fight is the worst outcome for both armies. Model this situation as an extensive game with perfect information and show that army 2 can increase its subgame perfect equilibrium payoff (and reduce army 1's payoff) by burning the bridge to its mainland (assume this act entails no cost), eliminating its option to retreat if attacked.

? EXERCISE 172.3 (Sharing heterogeneous objects) A group of n people have to share k objects that they value differently. Each person assigns values to the objects; no one assigns the same value to two different objects. Each person evaluates a set of objects according to the sum of the values she assigns to the objects in the set. The following procedure is used to share the objects. The players are ordered 1 through n. Person 1 chooses an object, then person 2 does so, and so on; if $k > n$, then after person n chooses an object, person 1 chooses a second object, then person 2 chooses a second object, and so on. Objects are chosen until none remain. (In Canada and the USA professional sports teams use a similar procedure to choose new players.) Denote by $G(n, k)$ the extensive game that models this procedure. If $k \leq n$ then obviously $G(n, k)$ has a subgame perfect equilibrium in which each player's strategy is to choose her favorite object among those remaining when her turn comes. Show that if $k > n$ then $G(n, k)$ may have no subgame perfect equilibrium in which person 1 chooses her favorite object on the first round. (You can give an example in which $n = 2$ and $k = 3$.) Now fix $n = 2$. Define x_k to be the object least preferred by the person who does *not* choose at stage k (i.e. who does not choose the last object); define x_{k-1} to be the object, among all those except x_k, least preferred by the person who does *not* choose at stage $k - 1$. Similarly, for any j with $2 \leq j \leq k$, given $x_j, \ldots, x_k$, define x_{j-1} to be the object, among all those excluding $\{x_j, \ldots, x_k\}$, least preferred by the person who does *not* choose at stage $j - 1$. Show that the game $G(2, 3)$ has a subgame perfect equilibrium in which for every $j = 1, \ldots, k$ the object x_j is chosen at stage j. (This result is true for $G(2, k)$ for all values of k.) If $n \geq 3$ then interestingly a person may be better off in all subgame perfect equilibria of $G(n, k)$ when she comes later in the ordering of players. (An example, however, is difficult to construct; one is given in Brams and Straffin 1979.)

The next exercise shows how backward induction can cause a relatively minor

change in the way in which a game ends to reverberate to the start of the game, leading to a very different action for the first-mover.

⁇ EXERCISE 173.1 (An entry game with a financially-constrained firm) An incumbent in an industry faces the possibility of entry by a challenger. First the challenger chooses whether or not to enter. If it does not enter, neither firm has any further action; the incumbent's payoff is TM (it obtains the profit M in each of the following $T \geq 1$ periods) and the challenger's payoff is 0. If the challenger enters, it pays the entry cost $f > 0$, and in each of T periods the incumbent first commits to fight or cooperate with the challenger in that period, then the challenger chooses whether to stay in the industry or to exit. (Note that the order of the firms' moves within a period differs from that in the game in Example 152.1.) If, in any period, the challenger stays in, each firm obtains in that period the profit $-F < 0$ if the incumbent fights and $C > \max\{F, f\}$ if it cooperates. If, in any period, the challenger exits, both firms obtain the profit zero in that period (regardless of the incumbent's action); the incumbent obtains the profit $M > 2C$ and the challenger the profit 0 in every subsequent period. Once the challenger exits, it cannot subsequently re-enter. Each firm cares about the sum of its profits.

a. Find the subgame perfect equilibria of the extensive game that models this situation.

b. Consider a variant of the situation, in which the challenger is constrained by its financial war chest, which allows it to survive at most $T - 2$ fights. Specifically, consider the game that differs from the one in part *a* in one respect: the histories in which (*i*) at the start of the game the challenger enters and (*ii*) the incumbent fights in $T - 1$ periods, are terminal histories (the challenger has to exit). For the terminal history in which the incumbent fights in the first $T - 1$ periods the incumbent's payoff is $M - (T - 2)F$ and the challenger's payoff is $-f - (T - 2)F$ (in period $T - 1$ the incumbent's payoff is 0, and in the last period its payoff is M). For the terminal history in which the incumbent cooperates in one of the first $T - 1$ periods and fights in the remainder of these periods and in the last period, the incumbent's payoff is $C - (T - 2)F$ and the challenger's payoff is $-f + C - (T - 2)F$. Find the subgame perfect equilibria of this game.

◆ EXAMPLE 173.2 (Dollar auction) Consider an auction in which an object is sold to the highest bidder, but *both* the highest bidder *and* the second highest bidder pay their bids to the auctioneer. When such an auction is conducted and the object is a dollar, the outcome is sometimes that the object is sold at a price *greater* than a dollar. (Shubik writes that "A total of payments between three and five dollars is not uncommon" (1971, 110).) Obviously such an outcome is inconsistent with a subgame perfect equilibrium of an extensive game that models the auction: every participant has the option of not bidding, so that in no subgame perfect equilibrium can anyone's payoff be negative.

Why, then, do such outcomes occur? Suppose that there are two participants,

and that both start bidding. If the player making the lower bid thinks that making a bid above the other player's bid will induce the other player to quit, she may be better off doing so than stopping bidding. For example, if the bids are currently \$0.50 and \$0.51, the player bidding \$0.50 is better off bidding \$0.52 *if* doing so induces the other bidder to quit, because she then wins the dollar and obtains a payoff of \$0.48, rather than losing \$0.50. The same logic applies even if the bids are greater than \$1.00, as long as they do not differ by more than \$1.00. If, for example, they are currently \$2.00 and \$2.01, then the player bidding \$2.00 loses only \$1.02 if a bid of \$2.02 induces her opponent to quit, whereas she loses \$2.00 if she herself quits. That is, in subgames in which bids have been made, the player making the second highest bid may optimally beat a bid that exceeds \$1.00, depending on the other players' strategies and the difference between the top two bids. (When discussing outcomes in which the total payment to the auctioneer exceeds \$1, Shubik remarks that "In playing this game, a large crowd is desirable ... the best time is during a party when spirits are high and the propensity to calculate does not settle in until at least two bids have been made" (1971, 109).)

In the next exercise you are asked to find the subgame perfect equilibria of an extensive game that models a simple example of such an auction.

? EXERCISE 174.1 (Dollar auction) An object that two people each value at v (a positive integer) is sold in an auction. In the auction, the people alternately have the opportunity to bid; a bid must be a positive integer greater than the previous bid. (In the situation that gives the game its name, v is 100 cents.) On her turn, a player may pass rather than bid, in which case the game ends and the other player receives the object; *both* players pay their last bids (if any). (If player 1 passes initially, for example, player 2 receives the object and makes no payment; if player 1 bids 1, player 2 bids 3, and then player 1 passes, player 2 obtains the object and pays 3, and player 1 pays 1.) Each person's wealth is w, which exceeds v; neither player may bid more than her wealth. For $v = 2$ and $w = 3$ model the auction as an extensive game and find its subgame perfect equilibria. (A much more ambitious project is to find all subgame perfect equilibria for arbitrary values of v and w.)

In all the extensive games studied so far in this chapter, each player has available finitely many actions whenever she moves. The next example shows how the procedure of backward induction may be used to find the subgame perfect equilibria of games in which a continuum of actions is available after some histories.

◆ EXAMPLE 174.2 (A synergistic relationship) Consider a variant of the situation in Example 37.2, in which two individuals are involved in a synergistic relationship. Suppose that the players choose their effort levels sequentially, rather than simultaneously. First individual 1 chooses her effort level a_1, then individual 2 chooses her effort level a_2. An effort level is a nonnegative number, and individual i's preferences (for $i = 1, 2$) are represented by the payoff function $a_i(c + a_j - a_i)$, where j is the other individual and $c > 0$ is a constant.

To find the subgame perfect equilibria, we first consider the subgames of length 1, in which individual 2 chooses a value of a_2. Individual 2's optimal action after the

history a_1 is her best response to a_1, which we found to be $\frac{1}{2}(c + a_1)$ in Example 37.2. Thus individual 2's strategy in any subgame perfect equilibrium is the function that associates with each history a_1 the action $\frac{1}{2}(c + a_1)$.

Now consider individual 1's action at the start of the game. Given individual 2's strategy, individual 1's payoff if she chooses a_1 is $a_1(c + \frac{1}{2}(c + a_1) - a_1)$, or $\frac{1}{2}a_1(3c - a_1)$. This function is a quadratic that is zero when $a_1 = 0$ and when $a_1 = 3c$, and reaches a maximum in between. Thus individual 1's optimal action at the start of the game is $a_1 = \frac{3}{2}c$.

We conclude that the game has a unique subgame perfect equilibrium, in which individual 1's strategy is $a_1 = \frac{3}{2}c$ and individual 2's strategy is the function that associates with each history a_1 the action $\frac{1}{2}(c + a_1)$. The outcome of the equilibrium is that individual 1 chooses $a_1 = \frac{3}{2}c$ and individual 2 chooses $a_2 = \frac{5}{4}c$.

? EXERCISE 175.1 (Firm–union bargaining) A firm's output is $L(100 - L)$ when it uses $L \leq 50$ units of labor, and 2500 when it uses $L > 50$ units of labor. The price of output is 1. A union that represents workers presents a wage demand (a nonnegative number w), which the firm either accepts or rejects. If the firm accepts the demand, it chooses the number L of workers to employ (which you should take to be a continuous variable, not an integer); if it rejects the demand, no production takes place ($L = 0$). The firm's preferences are represented by its profit; the union's preferences are represented by the value of wL.

a. Formulate this situation as an extensive game with perfect information.

b. Find the subgame perfect equilibrium (equilibria?) of the game.

c. Is there an outcome of the game that both parties prefer to any subgame perfect equilibrium outcome?

d. Find a Nash equilibrium for which the outcome differs from any subgame perfect equilibrium outcome.

? EXERCISE 175.2 (The "rotten kid theorem") A child's action a (a number) affects both her own private income $c(a)$ and her parent's income $p(a)$; for all values of a we have $c(a) < p(a)$. The child is selfish: she cares only about the amount of money she has. Her loving parent cares both about how much money she has and how much her child has. Specifically, her preferences are represented by a payoff equal to the smaller of the amount of money she has and the amount of money her child has. The parent may transfer money to the child. First the child takes an action, then the parent decides how much money to transfer. Model this situation as an extensive game and show that in a subgame perfect equilibrium the child takes an action that maximizes the sum of her private income and her parent's income. (In particular, the child's action does not maximize her own private income. The result is not limited to the specific form of the parent's preferences, but holds for any preferences with the property that a parent who is allocating a fixed amount x of money between herself and her child wishes to give more to the child when x is larger.)

? EXERCISE 176.1 (Comparing simultaneous and sequential games) The set of actions available to player 1 is A_1; the set available to player 2 is A_2. Player 1's preferences over pairs (a_1, a_2) are represented by the payoff $u_1(a_1, a_2)$, and player 2's preferences are represented by the payoff $u_2(a_1, a_2)$. Compare the Nash equilibria (in pure strategies) of the strategic game in which the players choose actions simultaneously with the subgame perfect equilibria of the extensive game in which player 1 chooses an action, then player 2 does so. (For each history a_1 in the extensive game, the set of actions available to player 2 is A_2.)

a. Show that if, for every value of a_1, a unique member of A_2 maximizes $u_2(a_1, a_2)$, then in every subgame perfect equilibrium of the extensive game, player 1's payoff is at least equal to her highest payoff in any Nash equilibrium of the strategic game.

b. Show that player 2's payoff in every subgame perfect equilibrium of the extensive game may be higher than her highest payoff in any Nash equilibrium of the strategic game.

c. Show that if for some values of a_1 more than one member of A_2 maximizes $u_2(a_1, a_2)$, then the extensive game may have a subgame perfect equilibrium in which player 1's payoff is less than her payoff in all Nash equilibria of the strategic game.

(For parts *b* and *c* you can give examples in which both A_1 and A_2 contain two actions. See Example 320.3 for further discussion of the implication of the order of play.)

TICKTACKTOE, CHESS, AND RELATED GAMES

Ticktacktoe, chess, and related games may be modeled as extensive games with perfect information. (A history is a sequence of moves and each player prefers to win than to tie than to lose.) Both ticktacktoe and chess may be modeled as finite games, so by Proposition 171.1 each game has a subgame perfect equilibrium. (The official rules of chess allow indefinitely long sequences of moves, but the game seems to be well modeled by an extensive game in which a draw is declared automatically if a position is repeated three times, rather than a player having the option of declaring a draw in this case, as in the official rules.) The subgame perfect equilibria of ticktacktoe are of course known, whereas those of chess are not (yet).

Ticktacktoe and chess are "strictly competitive" games (Definition 365.4): in every outcome, either one player loses and the other wins, or the players draw. A result in a later chapter implies that for such a game all Nash equilibria yield the same outcome (Corollary 369.1). Further, a player's Nash equilibrium strategy yields *at least* her equilibrium payoff, regardless of the other players' strategies (Proposition 367.3a). (The same is definitely not true for an arbitrary game that is not strictly competitive: look, for example, at the game in Figure 29.1.) Because any

subgame perfect equilibrium is a Nash equilibrium, the same is true for subgame perfect equilibrium strategies.

We conclude that in ticktacktoe and chess, either (*a*) one of the players has a strategy that guarantees she wins, or (*b*) each player has a strategy that guarantees at worst a draw.

In ticktacktoe, of course, we know that (*b*) is true. Chess is more subtle. In particular, it is not known whether White has a strategy that guarantees it wins, or Black has a strategy that guarantees it wins, or each player has a strategy that guarantees at worst a draw. The empirical evidence suggests that Black does not have a winning strategy, but this result has not been proved. When will a subgame perfect equilibrium of chess be found? (The answer "never" underestimates human ingenuity!)

? EXERCISE 177.1 (Subgame perfect equilibria of ticktacktoe) Ticktacktoe has subgame perfect equilibria in which the first player puts her first X in a corner. The second player's move is the same in all these equilibria. What is it?

? EXERCISE 177.2 (Toetacktick) Toetacktick is a variant of ticktacktoe in which a player who puts three marks in a line *loses* (rather than wins). Find a strategy of the first-mover that guarantees that she does not lose. (If fact, in all subgame perfect equilibria the game is a draw.)

? EXERCISE 177.3 (Three Men's Morris, or Mill) The ancient game of "Three Men's Morris" is played on a ticktacktoe board. Each player has three counters. The players move alternately. On each of her first three turns, a player places a counter on an unoccupied square. On each subsequent move, a player may move one of her counters to an adjacent square (vertically or horizontally, but not diagonally). The first player whose counters are in a row (vertically, horizontally, or diagonally) wins. Find a subgame perfect equilibrium strategy of player 1, and the equilibrium outcome.

Notes

The notion of an extensive game is due to von Neumann and Morgenstern (1944). Kuhn (1950b, 1953) suggested the formulation described in this chapter. The description of an extensive game in terms of histories was suggested by Ariel Rubinstein. The notion of subgame perfect equilibrium is due to Selten (1965). Proposition 171.1 is due to Kuhn (1953). The interpretation of a strategy when a subgame perfect equilibrium is interpreted as the outcome of the players' reasoning about each others' rational actions is due to Rubinstein (1991). The principle of optimality in dynamic programming is discussed by Bellman (1957, 83), for example.

The procedure in Exercises 162.2 and 172.1 was first studied by Mueller (1978) and Moulin (1981). The idea in Exercise 172.2 goes back at least to Sun-tzu, who,

in *The art of warfare* (probably written between 500BC and 300BC), advises "in surrounding the enemy, leave him a way out; do not press an enemy that is cornered" (end of Ch. 7; see, for example, Sun-tzu (1993, 132)). (That is, if no bridge exists in the situation described in the exercise, army 1 should build one.) Schelling (1966, 45) quotes Sun-tzu and gives examples of the strategy's being used in antiquity. My formulation of the exercise comes from Tirole (1988, 316). The model in Exercise 172.3 is studied by Kohler and Chandrasekaran (1971) and Brams and Straffin (1979). The game in Exercise 173.1 is based on Benoît (1984, Section 1). The dollar auction (Exercise 174.1) was introduced into the literature by Shubik (1971). Some of its subgame perfect equilibria, for arbitrary values of v and w, are studied by O'Neill (1986) and Leininger (1989); see also Taylor (1995, Chs. 1 and 6). Poundstone (1992, 257–272) writes informally about the game and its possible applications. The result in Exercise 175.2 is due to Becker (1974); see also Bergstrom (1989). The first formal study of chess is Zermelo (1913); see Schwalbe and Walker (2000) for a discussion of this paper and related work. Exercises 177.1, 177.2, and 177.3 are taken from Gardner (1959, Ch. 4), which includes several other intriguing examples.