1 The game of chess

Chapter summary

In the opening chapter of this book, we use the well-known game of chess to illustrate the notions of *strategy* and *winning strategy*. We then prove one of the first results in game theory, due to John von Neumann: in the game of chess either White (the first mover) has a winning strategy, or Black (the second mover) has a winning strategy, or each player has a strategy guaranteeing at least a draw. This is an important and nontrivial result, especially in view of the fact that to date, it is not known which of the above three alternatives holds, let alone what the winning strategy is, if one exists.

In later chapters of the book, this result takes a more general form and is applied to a large class of games.

We begin with an exposition of the elementary ideas in noncooperative game theory, by analyzing the game of chess. Although the theory that we will develop in this chapter relates to that specific game, in later chapters it will be developed to apply to much more general situations.

1.1 Schematic description of the game

The game of chess is played by two players, traditionally referred to as White and Black. At the start of a match, each player has sixteen pieces arranged on the chessboard. White is granted the opening move, following which each player in turn moves pieces on the board, according to a set of fixed rules. A match has three possible outcomes:

- Victory for White, if White captures the Black King.
- Victory for Black, if Black captures the White King.
- A draw, if:
  1. it is Black's turn, but he has no possible legal moves available, and his King is not in check;
  2. it is White's turn, but he has no possible legal moves available, and his King is not in check;
  3. both players agree to declare a draw;
  4. a board position precludes victory for both sides;
  5. 50 consecutive turns have been played without a pawn having been moved and without the capture of any piece on the board, and the player whose turn it is requests that a draw be declared;

6. or if the same board position appears three times, and the player whose turn it is requests that a draw be declared.

1.2 Analysis and results

For the purposes of our analysis all we need to assume is that the game is finite, i.e., the number of possible turns is bounded (even if that bound is an astronomically large number). This does not apply, strictly speaking, to the game of chess, but since our lifetimes are finite, we can safely assume that every chess match is finite.

We will denote the set of all possible board positions in chess by X. A board position by definition includes the identity of each piece on the board, and the board square on which it is located.

A board position, however, does not provide full details on the sequence of moves that led to it: there may well be two or sequences of moves leading to the same board position. We therefore need to distinguish between a "board position" and a "game situation," which is defined as follows.

Definition 1.1 *A* game situation *(in the game of chess) is a finite sequence* $(x_0, x_1, x_2, \ldots, x_K)$ *of board positions in* X *satisfying*

1. x_0 *is the opening board position.*
2. *For each even integer* k, $0 \le k < K$, *going from board position* x_K *to* x_{K+1} *can be accomplished by a single legal move on the part of White.*
3. *For each odd integer* k, $0 \le k < K$, *going from board position* x_K *to* x_{K+1} *can be accomplished by a single legal move on the part of Black.*

We will denote the set of game situations by H.

Suppose that a player wishes to program a computer to play chess. The computer would need a plan of action that would tell it what to do in any given game situation that could arise. A full plan of action for behavior in a game is called a *strategy*.

Definition 1.2 *A* strategy *for White is a function* s_W *that associates every game situation* $(x_0, x_1, x_2, \ldots, x_K) \in H$, *where* K *is even, with a board position* x_{K+1}, *such that going from board position* x_K *to* x_{K+1} *can be accomplished by a single legal move on the part of White.*

Analogously, a strategy *for Black is a function* s_B *that associates every game situation* $(x_0, x_1, x_2, \ldots, x_K) \in H$, *where* K *is odd, with a board position* x_{K+1} *such that going from board position* x_K *to* x_{K+1} *can be accomplished by a single legal move on the part of Black.*

Any pair of strategies (s_W, s_B) determines an entire course of moves, as follows. In the opening move, White plays the move that leads to board position $x_1 = s_W(x_0)$. Black then plays the move leading to board position $x_2 = s_B(x_0, x_1)$, and so on. The succeeding board positions are determined by $x_{2K+1} = s_W(x_0, x_1, \ldots, x_{2K})$ and $x_{2K+2} = s_B(x_0, x_1, \ldots, x_{2K+1})$ for all $K = 0, 1, 2, \ldots$.

1.2 Analysis and results

An entire course of moves (from the opening move to the closing one) is termed a *play* of the game.

Every play of the game of chess ends in either a victory for White, a victory for Black, or a draw. A strategy for White is termed a *winning strategy* if it guarantees that White will win, no matter what strategy Black chooses.

Definition 1.3 *A strategy s_W is a* winning strategy *for White if for every strategy s_B of Black, the play of the game determined by the pair (s_W, s_B) ends in victory for White. A strategy s_W is a* strategy guaranteeing at least a draw *for White if for every strategy s_B of Black, the play of the game determined by the pair (s_W, s_B) ends in either a victory for White or a draw.*

If s_W is a winning strategy for White, then any White player (or even computer program) adopting that strategy is guaranteed to win, even if he faces the world's chess champion.

The concepts of "winning strategy" and "strategy guaranteeing at least a draw" for Black are defined analogously, in an obvious manner.

The next theorem follows from one of the earliest theorems ever published in game theory (see Theorem 3.13 on page 46).

Theorem 1.4 *In chess, one and only one of the following must be true:*

(i) White has a winning strategy.
(ii) Black has a winning strategy.
(iii) Each of the two players has a strategy guaranteeing at least a draw.

We emphasize that the theorem does not relate to a particular chess match, but to all chess matches. That is, suppose that alternative (i) is the true case, i.e., White has a winning strategy s_W. Then any person who is the White player and follows the prescriptions of that strategy will always win every chess match he ever plays, no matter who the opponent is. If, however, alternative (ii) is the true case, then Black has a winning strategy s_B, and any person who is the Black player and follows the prescriptions of that strategy will always win every chess match he ever plays, no matter who the opponent is. Finally, if alternative (iii) is the true case, then White has a strategy s_W guaranteeing at least a draw, and Black has a strategy s_B guaranteeing at least a draw. Any person who is the White player (or the Black player) and follows the prescriptions of s_W (or s_B, respectively) will always get at least a draw in every chess match he ever plays, no matter who the opponent is. Note that if alternative (i) holds, there may be more than one winning strategy, and similar statements can be made with regard to the other two alternatives.

So, given that one of the three alternatives *must* be true, which one is it? We do not know. If the day ever dawns in which a winning strategy for one of the players is discovered, or strategies guaranteeing at least a draw for each player are discovered, the game of chess will cease to be of interest. In the meantime, we can continue to enjoy the challenge of playing (or watching) a good chess match.

Despite the fact that we do not know which alternative is the true one, the theorem is significant, because a priori it might have been the case that none of the alternatives was possible; one could have postulated that no player could ever have a strategy *always* guaranteeing a victory, or at least a draw.

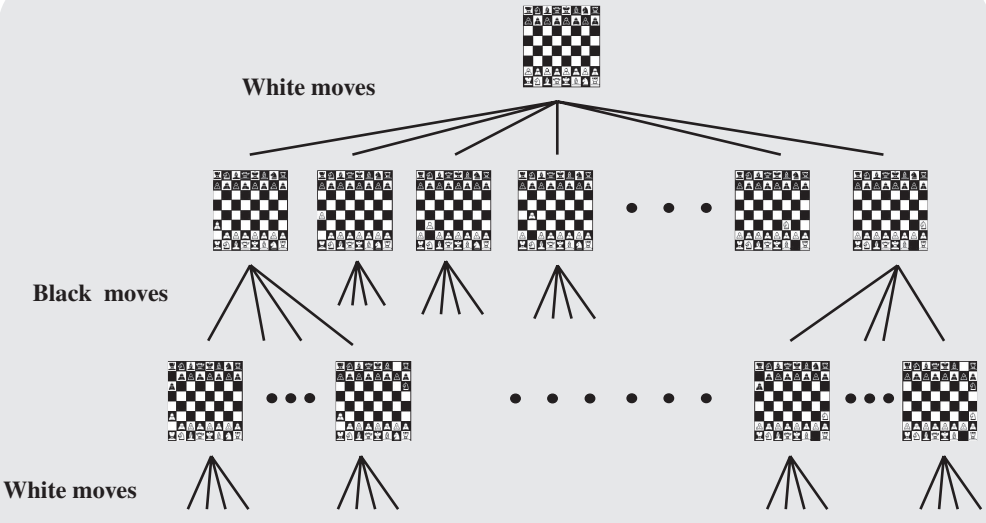

Figure 1.1 The game of chess presented in extensive form

We present two proofs of the theorem. The first proof is the "classic" proof, which in principle shows how to find a winning strategy for one of the players (if such a strategy exists) or a strategy guaranteeing at least a draw (if such a strategy exists). The second proof is shorter, but it cannot be used to find a winning strategy for one of the players (if such a strategy exists) or a strategy guaranteeing at least a draw (if such a strategy exists).

We start with several definitions that are needed for the first proof of the theorem. The set of game situations can be depicted by a tree[1] (see Figure 1.1). Such a tree is called a *game tree*. Each vertex of the game tree represents a possible game situation. Denote the set of vertices of the game tree by H.

The *root vertex* is the opening game situation x_0, and for each vertex x, the set of *children vertices* of x are the set of game situations that can be reached from x in one legal move. For example, in his opening move, White can move one of his pawns one or two squares forward, or one of his two rooks. So White has 20 possible opening moves, which means that the root vertex of the tree has 20 children vertices. Every vertex that can be reached from x by a sequence of moves is called a *descendant* of x. Every *leaf* of the tree corresponds to a terminal game situation, in which either White has won, Black has won, or a draw has been declared.

Given a vertex $x \in H$, we may consider the subtree beginning at x, which is by definition the tree whose root is x that is obtained by removing all vertices that are not descendants of x. This subtree of the game tree, which we will denote by $\Gamma(x)$, corresponds to a game that is called the *subgame beginning at* x. We will denote by n_x the number of vertices in $\Gamma(x)$. The game $\Gamma(x_0)$ is by definition the game that starts with the opening situation of the game, and is therefore the standard chess game.

1 The mathematical definition of a *tree* appears in the sequel (see Definition 3.5 on page 42).

1.2 Analysis and results

If y is a child vertex of x, then $\Gamma(y)$ is a subtree of $\Gamma(x)$ that does not contain x. In particular, $n_x > n_y$. Moreover, $n_x = 1$ if and only if x is a terminal situation of the game, i.e., the players cannot implement any moves at this subgame. In such a case, the strategy of a player is denoted by $\emptyset$.

Denote by

$$\mathcal{F} = \{\Gamma(x)\colon x \in H\} \tag{1.1}$$

the collection of all subgames that are defined by subtrees of the game of chess.

Theorem 1.4 can be proved using the result of Theorem 1.5.

Theorem 1.5 *Every game in $\mathcal{F}$ satisfies one and only one of the following alternatives:*

(i) White has a winning strategy.
(ii) Black has a winning strategy.
(iii) Each of the players has a strategy guaranteeing at least a draw.

Proof: The proof proceeds by induction on n_x, the number of vertices in the subgame $\Gamma(x)$.

Suppose x is a vertex such that $n_x = 1$. As noted above, that means that x is a terminal vertex. If the White King has been removed from the board, Black has won, in which case $\emptyset$ is a winning strategy for Black. If the Black King has been removed from the board, White has won, in which case $\emptyset$ is a winning strategy for White. Alternatively, if both Kings are on the board at the end of play, the game has ended in a draw, in which case $\emptyset$ is a strategy guaranteeing a draw for both Black and White.

Next, suppose that x is a vertex such that $n_x > 1$. Assume by induction that at all vertices y satisfying $n_y < n_x$, one and only one of the alternatives (i), (ii), or (iii) is true in the subgame $\Gamma(y)$.

Suppose, without loss of generality, that White has the first move in $\Gamma(x)$. Any board position y that can be reached from x satisfies $n_y < n_x$, and so the inductive assumption is true in the corresponding subgame $\Gamma(y)$. Denote by $C(x)$ the collection of vertices that can be reached from x in one of White's moves.

1. If there is a vertex $y_0 \in C(x)$ such that White has a winning strategy in $\Gamma(y_0)$, then alternative (i) is true in $\Gamma(x)$: the winning strategy for White in $\Gamma(x)$ is to choose as his first move the move leading to vertex y_0, and to follow the winning strategy in $\Gamma(y_0)$ at all subsequent moves.
2. If Black has a winning strategy in $\Gamma(y)$ for every vertex $y \in C(x)$, then alternative (ii) is true in $\Gamma(x)$: Black can win by ascertaining what the vertex y is after White's first move, and following his winning strategy in $\Gamma(y)$ at all subsequent moves.
3. Otherwise:
   - (1) does not hold, i.e., White has no winning strategy in $\Gamma(y)$ for any $y \in C(x)$. Because the induction hypothesis holds for every vertex $y \in C(x)$, either Black has a winning strategy in $\Gamma(y)$, or both players have a strategy guaranteeing at least a draw in $\Gamma(y)$.
   - (2) does not hold, i.e., there is a vertex $y_0 \in C(x)$ such that Black does not have a winning strategy in $\Gamma(y_0)$. But because (1) does not hold, White also does not have a

winning strategy in $\Gamma(y_0)$. Therefore, by the induction hypothesis applied to $\Gamma(y_0)$, both players have a strategy guaranteeing at least a draw in $\Gamma(y_0)$.

As we now show, in this case, in $\Gamma(x)$ both players have a strategy guaranteeing at least a draw. White can guarantee at least a draw by choosing a move leading to vertex y_0, and from there by following the strategy that guarantees at least a draw in $\Gamma(y_0)$. Black can guarantee at least a draw by ascertaining what the board position y is after White's first move, and at all subsequent moves in $\Gamma(y)$ either by following a winning strategy or following a strategy that guarantees at least a draw in that subgame. □

The proof just presented is a standard inductive proof over a tree: one assumes that the theorem is true for every subtree starting from the root vertex, and then shows that it is true for the entire tree. The proof can also be accomplished in the following way: select any vertex x that is neither a terminal vertex nor the root vertex. The subgame starting from this vertex, $\Gamma(x)$, contains at least two vertices, but fewer vertices than the original game (because it does not include the root vertex), and the induction hypothesis can therefore be applied to $\Gamma(x)$. Now "fold up" the subgame and replace it with a terminal vertex whose outcome is the outcome that is guaranteed by the induction hypothesis to be obtained in $\Gamma(x)$. This leads to a new game $\widehat{\Gamma}$. Since $\Gamma(x)$ has at least two vertices, $\widehat{\Gamma}$ has fewer vertices than the original game, and therefore by the induction hypothesis the theorem is true for $\widehat{\Gamma}$. It is straightforward to ascertain that a player has a winning strategy in $\widehat{\Gamma}$ if and only if he has a winning strategy in the original game.

In the proof of Theorem 1.5 we used the following properties of the game of chess:

(C1) The game is finite.
(C2) The strategies of the players determine the play of the game. In other words, there is no element of chance in the game; neither dice nor card draws are involved.
(C3) Each player, at each turn, knows the moves that were made at all previous stages of the game.

We will later see examples of games in which at least one of the above properties fails to hold, for which the statement of Theorem 1.5 also fails to hold (see for example the game "Matching Pennies," Example 3.20 on page 52).

We next present a second proof of Theorem 1.4. We will need the following two facts from formal logic for the proof. Let X be a finite set and let $A(x)$ be an arbitrary logical formula.[2] Then:

- If it is not the case that "for every $x \in X$ the formula $A(x)$ holds," then there exists an $x \in X$ where the formula $A(x)$ does not hold:
$$\neg(\forall x(A)) = \exists x(\neg A). \tag{1.2}$$
- If it is not the case that "there exists an $x \in X$ where the formula $A(x)$ holds," then for every $x \in X$ the formula $A(x)$ does not hold:
$$\neg(\exists x(A)) = \forall x(\neg A). \tag{1.3}$$

2 Recall that the logical statement "for every $x \in X$ event A obtains" is written formally as $\forall x(A)$, and the statement "there exists an $x \in X$ for which event A obtains" is written as $\exists x(A)$, while "event A does not obtain" is written as $\neg A$. For ease of exposition, we will omit the set X from each of the formal statements in the proof.

Second Proof of Theorem 1.4: As stated above, we assume that the game of chess is a finite game, i.e., there is a natural number K such that every play of the game concludes after at most $2K$ turns (K turns on the part of White and K turns on the part of Black). Assume that there are exactly $2K$ turns in every play of the game: every play that ends in fewer turns can be continued by adding more turns, up to $2K$, at which each player alternately implements the move "do nothing," which has no effect on the board position.

For every k, $1 \le k \le K$, denote by a_k the move implemented by White at his k-th turn, and by b_k the move implemented by Black at his k-th turn. Denote by W the sentence that White wins (after $2K$ turns). Then $\neg W$ is the sentence that the play ends in either a draw or a victory for Black. Using these symbols, the statement "White has a winning strategy" can be written formally as

$$\exists a_1 \forall b_1 \exists a_2 \forall b_2 \exists a_3 \cdots \exists a_K \forall b_K (W). \tag{1.4}$$

It follows that the statement "White does not have a winning strategy" can be written formally as

$$\neg(\exists a_1 \forall b_1 \exists a_2 \forall b_2 \exists a_3 \cdots \exists a_K \forall b_K (W)). \tag{1.5}$$

By repeated application of Equations (1.2) and (1.3) we deduce that this is equivalent to

$$\forall a_1 \exists b_1 \forall a_2 \exists b_2 \forall a_3 \cdots \forall a_K \exists b_K (\neg W). \tag{1.6}$$

This, however, says that Black has a strategy guaranteeing at least a draw. In other words, we have proved that if White has no winning strategy, then Black has a strategy that guarantees at least a draw. We can similarly prove that if Black has no winning strategy, then White has a strategy that guarantees at least a draw. This leads to the conclusion that one of the three alternatives of Theorem 1.4 must hold. □

1.3 Remarks

The second proof of Theorem 1.4 was brought to the attention of the authors by Abraham Neyman, to whom thanks are due.

1.4 Exercises

1.1 "The outcome of every play of the game of chess is either a victory for White, a victory for Black, or a draw." Is that statement equivalent to the result of Theorem 1.4? Justify your answer.

1.2 Find three more games that satisfy Properties (C1)–(C3) on page 6 that are needed for proving Theorem 1.4.

1.3 Theorem 1.4 was proved in this chapter under the assumption that the length of a game of chess is bounded. In this exercise we will prove the theorem without that assumption, that is, we will allow an infinite number of moves. We will agree that the outcome of an infinitely long game of chess is a draw.

When one allows infinite plays, the set of game situations is an infinite set. However, to know how to continue playing, the players need not know all the sequence of past moves. In fact, only a bounded amount of information needs to be told to the players, e.g.,

- What is the current board position?
- Have the players played an even or an odd number of moves up to now (for knowing whose turn it is)?
- For every board position, has it appeared in the play up to now 0 times, once, or more than once (for knowing whether the player whose turn it is may ask for a draw)?

We will therefore make use of the fact that one may suppose that there are only a finite number of board positions in chess.

Consider the following version of chess. The rules of the game are identical to the rules on page 1, with the one difference that if a board position is repeated during a play, the play ends in a draw. Since the number of game situations is finite, this version of chess is a finite game. We will call it "finite chess."

(a) Prove that in finite chess exactly one of the following holds:
  (i) White has a winning strategy.
  (ii) Black has a winning strategy.
  (iii) Each of the two players has a strategy guaranteeing at least a draw.

(b) Prove that if one of the players has a winning strategy in finite chess, then that player also has a winning strategy in chess.

We now prove that if each player has a strategy guaranteeing at least a draw in finite chess, then each player has a strategy guaranteeing at least a draw in chess. We will prove this claim for White. Suppose, therefore, that White has a strategy σ_W in finite chess that guarantees at least a draw. Consider the following strategy $\widehat{\sigma}_W$ for White in chess:

- Implement strategy σ_W until either the play of chess terminates or a board position repeats itself (at which point the play of finite chess terminates).
- If the play of chess arrives at a game situation x that has previously appeared, implement the strategy σ_W restricted to the subgame beginning at x until the play arrives at a board position y that has previously appeared, and so on.

(c) Prove that the strategy $\widehat{\sigma}_W$ guarantees at least a draw for White in chess.

2 Utility theory

Chapter summary

The objective of this chapter is to provide a quantitative representation of players' preference relations over the possible outcomes of the game, by what is called a *utility function*. This is a fundamental element of game theory, economic theory, and decision theory in general, since it facilitates the application of mathematical tools in analyzing game situations whose outcomes may vary in their nature, and often be uncertain.

The utility function representation of preference relations over uncertain outcomes was developed and named after John von Neumann and Oskar Morgenstern. The main feature of the von Neumann–Morgenstern utility is that it is linear in the probabilities of the outcomes. This implies that a player evaluates an uncertain outcome by its *expected utility*.

We present some properties (also known as axioms) that players' preference relations can satisfy. We then prove that any preference relation having these properties can be represented by a von Neumann–Morgenstern utility and that this representation is determined up to a positive affine transformation. Finally we note how a player's attitude toward risk is expressed in his von Neumann–Morgenstern utility function.

2.1 Preference relations and their representation

A game is a mathematical model of a situation of interactive decision making, in which every decision maker (or *player*) strives to attain his "best possible" outcome, knowing that each of the other players is striving to do the same thing.

But what does a player's "best possible" outcome mean? The outcomes of a game need not be restricted to "Win," "Loss," or "Draw." They may well be monetary payoffs or non-monetary payoffs, such as "your team has won the competition," "congratulations, you're a father," "you have a headache," or "you have granted much-needed assistance to a friend in distress."

To analyze the behavior of players in a game, we first need to ascertain the set of outcomes of a game and then we need to know the preferences of each player with respect to the set of outcomes. This means that for every pair of outcomes x and y, we need to know for each player whether he prefers x to y, whether he prefers y to x, or whether he is indifferent between them. We denote by O the set of outcomes of the game. The preferences of each player over the set O are captured by the mathematical concept that is termed *preference relation*.

Definition 2.1 *A* preference relation *of player i over a set of outcomes O is a binary relation denoted by $\succsim_i$.*

A binary relation is formally a subset of $O \times O$, but instead of writing $(x, y) \in \succsim_i$ we write $x \succsim_i y$, and read that as saying "player i either prefers x to y or is indifferent between the two outcomes"; sometimes we will also say in this case that the player "weakly prefers" x to y. Given the preference relation $\succsim_i$ we can define the corresponding *strict preference relation* $\succ_i$, which describes when player i strictly prefers one outcome to another:

$$x \succ_i y \iff x \succsim_i y \text{ and } y \not\succsim_i x. \tag{2.1}$$

We can similarly define the *indifference* relation $\approx_i$, which expresses the fact that a player is indifferent between two possible outcomes:

$$x \approx_i y \iff x \succsim_i y \text{ and } y \succsim_i x. \tag{2.2}$$

We will assume that every player's preference relation satisfies the following three properties.

Assumption 2.2 *The preference relation $\succsim_i$ over O is* complete*; that is, for any pair of outcomes x and y in O either $x \succsim_i y$, or $y \succsim_i x$, or both.*

Assumption 2.3 *The preference relation $\succsim_i$ over O is* reflexive*; that is, $x \succsim_i x$ for every $x \in O$.*

Assumption 2.4 *The preference relation $\succsim_i$ over O is* transitive*; that is, for any triple of outcomes x, y, and z in O, if $x \succsim_i y$ and $y \succsim_i z$ then $x \succsim_i z$.*

The assumption of completeness says that a player should be able to compare any two possible outcomes and state whether he is indifferent between the two, or has a definite preference for one of them, in which case he should be able to state which is the preferred outcome. One can imagine real-life situations in which this assumption does not obtain, where a player is unable to rank his preferences between two or more outcomes (or is uninterested in doing so). The assumption of completeness is necessary for the mathematical analysis conducted in this chapter.

The assumption of reflexivity is quite natural: every outcome is weakly preferred to itself.

The assumption of transitivity is needed under any reasonable interpretation of what a preference relation means. If this assumption does not obtain, then there exist three outcomes x, y, z such that $x \succsim_i y$ and $y \succsim_i z$, but $z \succ_i x$. That would mean that if a player were asked to choose directly between x and z he would choose z, but if he were first asked to choose between z and y and then between the outcome he just preferred (y) and x, he would choose x, so that his choices would depend on the order in which alternatives are offered to him. Without the assumption of transitivity, it is unclear what a player means when he says that he prefers z to x.

The greater than or equal to relation over the real numbers $\geq$ is a familiar preference relation. It is complete and transitive. If a game's outcomes for player i are sums of dollars, it is reasonable to suppose that the player will compare different outcomes using this preference relation. Since using real numbers and the $\geq$ ordering relation is very convenient for the purposes of conducting analysis, it would be an advantage to be able