Subgame Perfect Nash Equilibrium

IE 5545: Decision Analysis
Prof. Krishnamurthy Iyer*

May 2, 2026

Having described the representation of dynamic games of complete information, we now consider the question of analyzing the behavior of players in such a game. Our goal is to develop a notion of equilibrium that we expect rational players to follow in such games. To motivate this, we begin by considering a very simple example.

1 Sequential coordination game

Consider the sequential version of the coordination game analyzed earlier, where Alice and Bob have to decide between two options: whether to attend a Salsa-dancing workshop (S) or attend the Star Trek convention (ST). In this sequential version, we assume that Alice makes her choice first, Bob observes the choice and then makes his decision. The game tree is as in Figure 1.

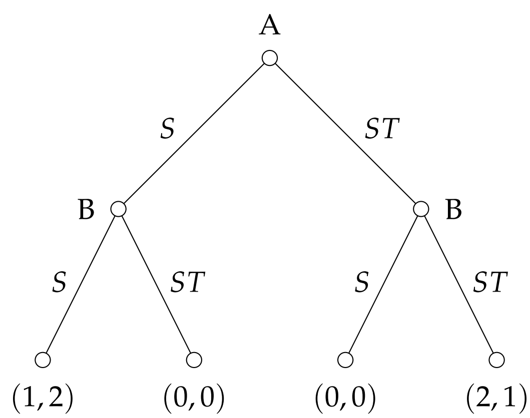

Figure 1: Sequential coordination game.

In this game, Alice has two pure strategies S and ST, whereas Bob has four pure strategies (S, S), (S, ST), (ST, S) and (ST, ST), where for a pure strategy (x, y), Bob plays x after Alice plays S and plays y after Alice plays ST. Since we know the set of pure strategies, we can write down the strategic form for the game, as in Figure 2.

Notice that moving from the extensive form in Figure 1 to the strategic form in Figure 2 destroys all the timing information embedded in the game tree. Specifically, in the strategic form, we assume that Alice and Bob choose their strategies *simultaneously*, and play accordingly. So, under the

	B				
		(S,S)	(S,ST)	(ST,S)	(ST,ST)
A	S	(1,2)	(1,2)	(0,0)	(0,0)
	ST	(0,0)	(2,1)	(0,0)	(2,1)

Figure 2: Strategic form of sequential coordination game

strategy profile $(S,(ST,ST))$, Alice chooses strategy S and Bob chooses (ST,ST) simultaneously, and once this choice is made, they implement their respective strategy.

One approach we can take to analyze the original sequential game, is to consider its strategic form, and identify its (pure or mixed) Nash equilibria, and claim that these capture how rational players will plan in the sequential game. However, as mentioned above, this process destroys the timing information, and an important question is whether this destructive process fails to capture some critical aspect of the dynamic game.

To investigate this question, let us analyze the strategic form of the sequential coordination game (Figure 2), and compute its set of (pure) Nash equilibria. Doing so reveals that this dynamic game has three pure Nash equilibria:

1. $(S,(S,S))$;
2. $(ST,(S,ST))$ and;
3. $(ST,(ST,ST))$.

Thus, in the $(S,(S,S))$ Nash equilibrium, Bob decides to go to Salsa-dancing workshop ((S,S)) no matter what Alice decides, and Alice knowing this, decides to go to Salsa dancing workshop S. Both players are indeed playing best responses given the other player's strategy.

However, we see that the $(S,(S,S))$ equilibrium is somewhat unsatisfying, because this requires Alice to believe that Bob will choose S even if she chooses ST. However, given Alice chooses ST, it is not rational for Bob to choose S, since choosing S gets Bob a payoff of 0, whereas choosing ST gets him a payoff of 1. Hence, this belief of Alice cannot be *justified* unless Alice thinks Bob is not rational (or unless she has some uncertainty about Bob's payoffs). Since we make the assumption that the rationality of players and the structure of the game is common knowledge, we conclude that Alice is not justified in believing that Bob will play S if she plays ST.

Another way to look at the $(S,(S,S))$ equilibrium, is through the notion of "cheap talk." Suppose Alice and Bob are allowed to talk to each other before they play the sequential coordination game. (Such talk is referred to as cheap talk, since players are not committed to follow through on what they discuss.) Then, one way such an equilibrium could arise is if Bob "threatens" Alice that he will choose S no matter what Alice chooses. Alice then *believes* this threat and chooses S. However, note that Bob is making a threat of playing a sub-optimal action, and if Alice knows that Bob is rational, then she should not believe this threat. In other words, Bob is making a threat that is not credible — if Alice is rational, then she can just call Bob's bluff and choose ST, in which case, the only optimal action for Bob is to play ST.

Similarly, in the $(ST,(ST,ST))$ equilibrium, Bob is making the threat that he will choose ST even if Alice chooses S. Again, this is a non-credible threat, because Alice can easily call Bob's bluff by choosing S, in which case Bob will choose S. (In this case, Alice may not want to call Bob's bluff; nevertheless it is still a non-credible threat.) Again, since this equilibrium involves non-credible threat, we find this equilibrium unsatisfying.

On the other hand, in the $(ST, (S, ST))$ equilibrium, Bob is making the "threat" (or a "promise") that he will play S is Alice plays S and will play ST if Alice plays ST. This is a credible threat/promise, since Bob is only saying that he will respond optimally to Alice's choice. Alice is completely justified in believing this threat, and consequently, she ends up playing ST. In this equilibrium, the play is consistent with the common knowledge of the players' rationality.

From this discussion, we see that ignoring the timing information embedded in the extensive form may lead to Nash equilibria that involve some players making non-credible threats, and other players believing such non-credible threats. Since rationality of players is common-knowledge, we want to avoid such equilibrium. In other words, we would like to require that in equilibrium no rational player ever believes a non-credible threat (and consequently, no rational player ever makes one).

How can we find a Nash equilibrium that does not involve non-credible threats? In the sequential coordination game, the equilibrium $(ST, (S, ST))$ can be found by "working backwards": the only optimal action for Bob after Alice plays S is to play S, and the only optimal action for Bob after Alice plays ST is to play ST. This completely determines Bob's strategy. Given this strategy of Bob, the only best response for Alice is to play ST. Our goal is generalize this process to all dynamic games, even those with imperfect information.

2 Subgames perfectness

The process of "working backwards", known as *backward induction*, is easy to apply in games of perfect information. Essentially, we first consider each decision node that is only connected to terminal nodes. For these decision nodes, we determine the optimal action for the associated player. Next, we consider all decision nodes that are only connected to terminal nodes or the decision nodes already considered. Using the optimal actions that have been already computed, we determine the optimal actions at these nodes for the associated players. We continue this procedure, where at each step we consider only those decision nodes that connect to either terminal nodes or the decision nodes that have been already analyzed, and for these decision nodes we determine the optimal action. We stop when we reach the root. Then, we put together all the optimal actions to determine the Nash equilibrium. Note that since backward induction ensures that each player is playing an optimal action at each of her decision node (even ones that won't be played), we obtain that no player makes non-credible threat, and no player believes non-credible threat. Thus, in games of perfect information, backward induction will always give us an Nash equilibrium of the type we want.

However, this backward induction procedure fails if we have a game of imperfect information, since a player cannot choose different actions at two decision nodes that belong to the same information set. Essentially, we cannot split the analysis of all the decision nodes that belong to the same information set. Instead, we must analyze all these decision nodes simultaneously. More generally, in a game of imperfect information, certain "parts" of the game may have to be analyzed as a whole — splitting them further may change the game completely. A "part" of a game that can be analyzed as a whole constitutes a *subgame*, defined as follows.

Definition 1 (Subgame)**.** *Given an extensive form E, a subgame G of E consists of a single decision node x and all its descendants in E such that the following conditions hold:*

1. *the decision node x is in a singleton information set. In other words, there is no decision node y such that x and y are in the same information set;*

2. *No information set is broken while including the descendants of x.*

Note that by this definition E itself is a subgame of E. But there might be other subgames of E. A proper subgame G of E is any subgame $G \neq E$.

For example, the sequential coordination game has three subgames — one subgame which is the entire game, the second subgame following Alice's choice of S, and a third subgame following Alice's choice of ST. More generally, we give some examples (and non-examples) of subgames in Figure 3.

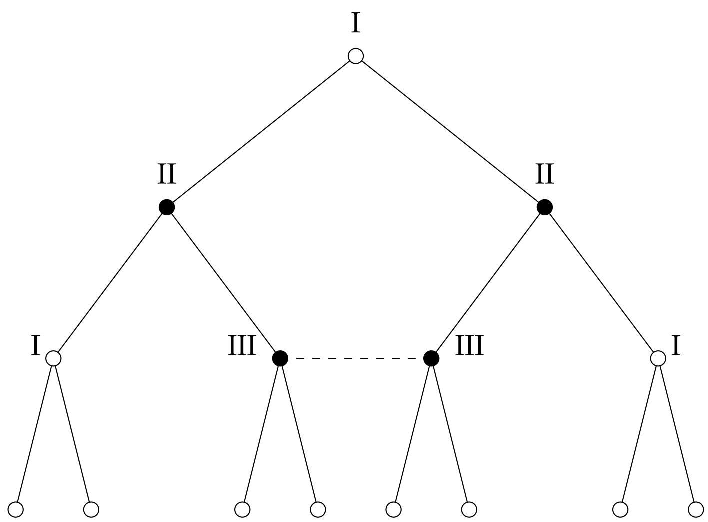

Figure 3: Examples of subgames: Here, for each white node, the node and its descendants form a subgame. On the other hand, for each black node, the node and its descendants do not form a subgame. For each black node, the node and its descendants necessarily break at least one information set.

A natural generalization of backward induction to games of imperfect information is to require that in each subgame, players are playing rationally. This is required of even those subgames that will not be played. Note that this is indeed true in the sequential coordination game in the $(ST, (S, ST))$ equilibrium, but not in the other two pure NE. This leads us to a *refinement* of the notion of Nash equilibrium, as follows:

Definition 2 (Subgame Perfect Nash Equilibrium (SPNE))**.** *In a dynamic game of complete information E, a strategy profile $\sigma = (\sigma_1, \cdots, \sigma_N)$ is an SPNE if the strategy profile induces a Nash equilibrium in every subgame of E, even those that are not played.*

Note that since E itself is a subgame of E, we require that an SPNE σ must also be an NE of E. In other words, all SPNE of E are NEs. On the other hand, there might exist NEs of E that are not SPNEs. (For example, in coordination game, $(S, (S, S))$ is an NE, but not an SPNE, because the strategy profile $(S, (S, S))$ does not induce an NE in the subgame starting at the decision node after Alice chooses ST.)

We can extend the process of backward induction to find an SPNE for any finite extensive form game E as follows.

Let $E_0 = E$ and $t = 0$. We perform the following steps until we reach stop:

1. We first determine all subgames $G \subseteq E_t$ that do not contain any proper subgames themselves. For any such subgame G, we write down its strategic form, and find a Nash equilibrium. (If G has multiple NE, we choose one arbitrarily.)

   If this step already determines an NE for E_t, we stop.

2. Otherwise, we replace each G by a terminal node with payoffs corresponding to that under the chosen NE. At the end of this step, we obtain a new extensive form game E_{t+1}. We then proceed to step 1 with $t \leftarrow t+1$.

Since E is a finite extensive form game, each subgame (and its strategic form) is finite, and thus has a Nash equilibrium (by Nash's existence theorem). Thus, we can perform each step in the preceding algorithm.

Moreover, at each iteration, we remove at least one proper subgame from E_t, and hence $E_{t+1} \subsetneq E_t$. Thus, the above process will eventually stop. When that happens, we will have an NE in each subgame of E. In particular, we will have determined an action at each information set I_i of player i. We then "stitch together" these actions to obtain a strategy σ_i for each player i. The resulting strategy profile $\sigma = (\sigma_1, \ldots, \sigma_N)$ ensures that in each subgame the players are playing a Nash equilibrium, and hence constitutes a subgame perfect Nash equilibrium of E.

Since the above algorithm produces an SPNE for any finite extensive form game, it also proves that every finite extensive form game has at least one SPNE. There can be more than one, if any subgame has more than one NE.