Iterated Backward Inference: An Algorithm for Proper Rationalizability

Oliver Schulte
School of Computing Science and
Department of Philosophy
Simon Fraser University
Vancouver, Canada
oschulte@cs.sfu.ca

May 18, 2003

Abstract

An important approach to game theory is to examine the consequences of beliefs that agents may have about each other. This paper investigates *respect for public preferences.* Consider an agent A who believes that B strictly prefers an option a to an option b. Then A *respects* B's preference if A assigns probability 1 to the choice of a given that B chooses a or b. Respect for *public* preferences requires that if it is common belief that B prefers a to b, then it is common belief that all other agents respect that preference. Along the lines of Blume, Brandenburger and Dekel [3] and Asheim [1], I treat respect for public preferences as a constraint on lexicographic probability systems. The main result is that given respect for public preferences and perfect recall, players choose in accordance with Iterated Backward Inference. Iterated Backward Inference is a procedure that generalizes standard backward induction reasoning for games of both perfect and imperfect information. From Asheim's characterization of proper rationalizability [1] it follows that properly rationalizable strategies are consistent with respect for public preferences; hence strategies eliminated by Iterated Backward Inference are not properly rationalizable.

1 Introduction and Overview

Game theory provides a general formalism for representing strategic interactions. The question arises what we can predict about the behaviour of the agents in a given situation. A *solution concept* gives a formal answer to this question, by associating a set of game plays—the "solution set"—with a given game matrix or game tree. An important line of research examines epistemic assumptions that validate a given solution concept (cf. [14]). What conditions imply that the predictions of the solution concept are correct?

This paper examines the implications of *respect for public preferences.* Blume *et al.* introduced the concept of respect for preferences [3, Def. 4] to characterize Myerson's "proper equilibrium" [8]. Let us assume that each agent's choices are based on a lexicographic probability system (LPS) $\rho = (\rho^1, .., \rho^k)$, where each ρ^i is a probability measure over the choices of other agents. According to Blume, Brandenburger, and Dekel, "the first component of the LPS [i.e., ρ^1] can be thought of as representing the player's primary theory..." [3, page 82]. In a lexicographic probability system, the probability of an event E may be defined even conditional on an event E' that the agent believes not to obtain. If ρ_B is the LPS of agent B, then for any choice of an agent A between two options a and b, we may

consider the conditional LPS $\rho_B|\{a,b\}$. According to Blume *et al.*, B respects the preferences of A if $[\rho_B^1|\{a,b\}](a) = 1$ whenever A strictly prefers option a to b. Intuitively, the "primary hypothesis" of B is that A chooses a, given that A chooses either a or b and prefers a. For example, suppose that agent A has three options, \$300, \$200, \$100. Then if A prefers \$200 to \$100, respect for preferences requires that $[\rho_B^1|\{\$200, \$100\}](\$200) = 1$.

Asheim [1] introduced a weaker condition: according to his definition, B respects the preferences of A if $[\rho_B^1|\{a,b\}](a) = 1$ whenever B *believes* (with certainty - see Section 4) that A strictly prefers option a to b. We may think of the definition of Blume *et al.* as a special case where B's beliefs about A's preferences are true—as they well may be at equilibrium.

Respect for public preferences requires common belief that $[\rho_B^1|\{a,b\}](a) = 1$ whenever it is *common belief* among all agents that A strictly prefers option a to b. An event is common belief among the agents if all agents believe that it obtains, all agents believe all agents believe that it obtains, etc. Preferences that are common belief are "public" in the sense that all agents are aware of them. Assuming that agents know their own preferences, then if A believes that she prefers a to b, this is indeed the case, and hence public preferences are true preferences.

This paper is a formal investigation of what respect for public preferences implies about agents' behaviour. More specifically, I derive consequences of the following assumptions:

Respect for Public Preferences If it is common belief that player A strictly prefers option a over b, then it is common belief that $[\rho_B^1|\{a,b\}](a) = 1$ for each player $B \neq A$.

Full Lexicographic Rationality It is common belief that each player maximizes lexicographic expected utility, with respect to an LPS with *full support*. A lexicographic probability system ρ has full support if every nonempty event receives positive probability at some measure in ρ.

I specify an iterated elimination procedure that computes consequences of these assumptions, which I term Iterated Backward Inference (IBI). In two special cases, IBI coincides with other well-known algorithms. First, in a game of perfect information with a unique backward induction solution, the result of IBI is that solution. Second, suppose we represent a strategic form game as a game tree with two information sets, one for each player with moves corresponding to strategies. Then IBI coincides with the *Dekel-Fudenberg* procedure [4]: First, eliminate all weakly dominated strategies. Then iteratively eliminate all strictly dominated strategies. Thus IBI generalizes at once both standard backward induction and the Dekel-Fudenberg procedure.

Applying Asheim's characterization of Schuhmacher's concept of proper rationalizability [1], [11], it is easy to show that properly rationalizable strategies are consistent with Respect for Public Preferences. Hence if IBI eliminates a strategy s_i, then s_i is not properly rationalizable. So IBI can be used to find strategies that are not properly rationalizable.

The paper is organized as follows. Sections 2 and 3 define standard game-theoretic notions such as game trees and strategies, and review definitions pertaining to lexicographic probability systems. Sections 4 and 5 formalize a number of epistemic assumptions, particularly Respect for Public Preferences. The remainder of the paper investigates the consequences of these assumptions. Sections 6 and 7 define Iterated Backward Inference, establish its correctness and show existence for finite games—that is, in finite games some strategy profile is guaranteed to survive IBI.

2 Preliminaries: Game Trees and Strategies

This section gives the standard definition of *sequential games* or *game trees*. A tree V is a directed graph with a root node r such that for every node v in the tree, there is exactly one path from r to v.

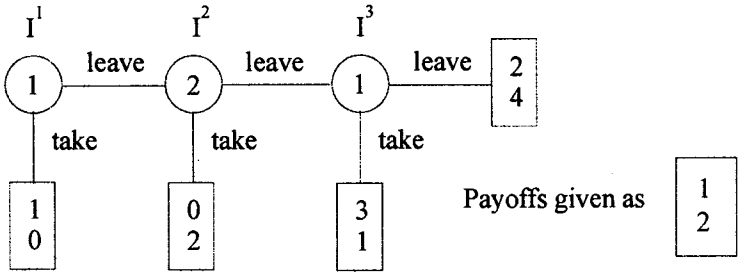

Figure 1: The 3-move version of the Centipede game with players A and B.

I write $x > x'$ to indicate that node x' is a successor of x. The edges in the graph are labelled with "actions". The set of actions available at a node x is denoted by $A(x)$. Nodes in the tree correspond to sequences of actions. The definition of a finite game tree is as follows.

Definition 1 *A finite* ***sequential game*** T *is a tuple* $\langle N, (X, E), player, \{I_i\}, \{u_i\}\rangle$ *whose components are the following.*

1. *A finite set* N *(the set of* ***players****).*
2. *A finite tree* (X, E) *labelled with actions.*
3. *A function* player *that assigns to each nonleaf node in* X *a member of* N*.*
4. *For each player* $i \in N$ *an* ***information partition*** $\mathcal{I}_i$ *defined on* $\{x \in X : player(x) = i\}$*. An element* I_i *of* $\mathcal{I}_i$ *is called an* ***information set*** *of player* i*. We require that if* x, x' *are members of the same information set* I_i*, then* $A(x) = A(x')$*. Let* $A_i =_{df} \cup\{A(x) : x \in X \text{ and } player(x) = i\}$ *denote the set of actions of player* i*.*
5. *For each player* $i \in N$ *a* ***payoff function*** $u_i : Z \to R$ *that assigns a real number to each leaf node.*

Though the general notion of an extensive form game permits "chance" moves by "nature", Definition 1 does not include chance moves. Another simplification that I make throughout the paper is to consider only 2-player games, that is, I take $N = \{1, 2\}$. It is straightforward to generalize the results in this paper to games with chance moves and any finite number of players, but doing so gives rise to technical complications that do not illuminate the main issues. I illustrate Definition 1 in two extensive form games, the three-move version of the well-known Centipede Game (Figure 1) and a game due to Kohlberg (Figure 2) [7]. There are 7 nodes in the Centipede Game. The terminal nodes comprise *leave* $*$ *leave* $*$ *leave* and all sequences ending in *take* ($*$ denotes concatenation). Each of the three information sets I^1, I^2, I^3 is a singleton, which makes the Centipede game a game of **perfect information**. Kohlberg's game is a game of imperfect information because I^4 contains two nodes.

If i denotes a player, I write $-i$ for the opponent of player i; so $-1 = 2$ and $-2 = 1$ when $1, 2$ refer to players. A **strategy** for player i is a function $s_i : \{x \in H : player(x) = i\} \to A_i$, such that (1) $s_i(x) \in A(x)$ for all nodes x belonging to player i, and (2) if $I(x) = I(x')$, then $s_i(x) = s_i(x')$. I write $S_i(T)$ for the **set of strategies** of player i in T. A strategy pair (s_1, s_2) of players 1 and 2 respectively determines a unique terminal history denoted by $play(s_1, s_2)$. I extend the utility functions u_i to strategy pairs by defining $u_i(s_1, s_2) =_{df} u_i(play(s_1, s_2))$. I assume throughout the paper that

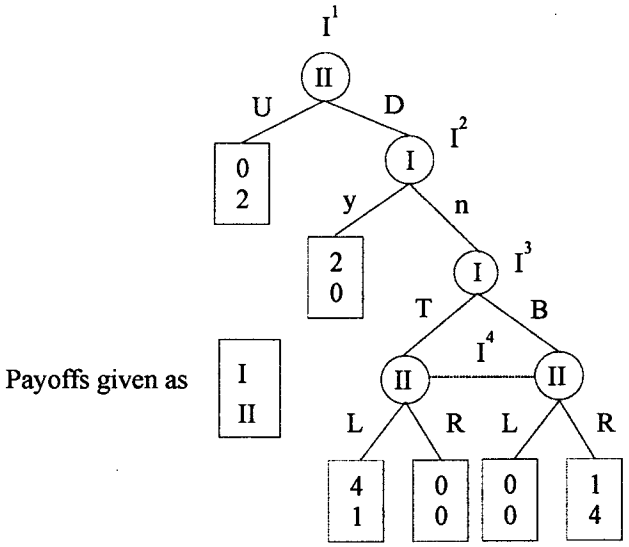

Figure 2: An extensive form game due to Kohlberg

1/2	l_2	t_2
tt	1,0	1,0
tl	1,0	1,0
lt	3,1	0,2
ll	2,4	0,2

Table 1: The strategic form of the Centipede Game

every node is reachable by some pair of strategies. That is, I assume that for all nodes $x \in H$, there is a strategy pair (s_1, s_2) such that x is reached along $play(s_1, s_2)$ (cf. [6, Sec. 2.1]).

To illustrate, in the Centipede Game there are two strategies for player 2, which I denote by l_2 and t_2 where $l_2(\emptyset * leave) = leave$, and $t_2(\emptyset * leave) = take$. Thus if T denotes the Centipede Game, then $S_2(T) = \{l_2, t_2\}$. Player 1 has four strategies, specifying choices at I^1 and I^3. We may use the tuples in $\{l, t\} \times \{l, t\}$ to denote these strategies, such that for example $lt(\emptyset) = leave$, and $lt(\emptyset * leave * leave) = take$. The play sequence resulting from the strategy pair (lt, t_2) is $\emptyset * leave * take$; in our notation $play(lt, t_2) = \emptyset * leave * take$. Thus $u_1(lt, t_2) = 0$, and $u_2(lt, t_2) = 2$. A matrix whose rows correspond to strategies for player 1 and columns to strategies for player 2 gives the **normal form**, or **strategic form**, of a game tree. Table 1 shows the strategic form of the Centipede game. In Kohlberg's game, each player has four strategies. Table 2 shows the normal form of Kohlberg's game.

I/II	UL	UR	DL	DR
yT	0,2	0,2	2,0	2,0
yB	0,2	0,2	2,0	2,0
nT	0,2	0,2	4,1	0,0
nB	0,2	0,2	0,0	1,4

Table 2: The strategic form of Kohlberg's game

ρ_2^j/X	tt	tl	lt	ll
ρ_2^3	0	0	1/2	1/2
ρ_2^2	0	0	1	0
ρ_2^1	1/2	1/2	0	0

Table 3: A lexicographic probability system ρ_2 representing the beliefs of player 2 about the strategies of player 1 in the Centipede Game

A crucial question in the developments below is how player i ranks the strategies of her opponent $-i$ given the information in some information set I. To describe this formally, I map information sets into sets of strategies as follows. First, define $[I] = \{(s_1, s_2) : play(s_1, s_2)$ intersects $I\}$. With respect to player i's uncertainty space S_{-i}, the information in I corresponds to the set of strategies $[I]_{-i}$ of player $-i$ that are consistent with I; formally I define $[I]_{-i} = \{s_{-i} : \exists s_i.(s_i, s_{-i}) \in [I]\}$. To illustrate, in the Centipede game of Figure 1, $[I^2] = \{(ll, l_2), (lt, l_2), (ll, t_2), (lt, t_2)\}$, so $[I^2]_1 = \{ll, lt\}$ and $[I^2]_2 = \{l_2, t_2\}$.

Following [6, Sec. 2.1], I define $I > I'$ iff there is $x \in I, x' \in I'$ such that $x > x'$, and $I \geq I'$ iff $I > I'$ or $I = I'$. Thus $I > I'$ holds if some play sequence reaches I' and then I. For example, in the Kohlberg game of Figure 2, we have that $I^j > I^4$ for all information sets $I^j \neq I^4$.

3 Preliminaries: Lexicographic Expected Utility

Let X be a finite set of points. A **lexicographic probability system** over X is a finite sequence $\rho = (\rho^1, \rho^2, ..., \rho^k)$, where each ρ^j is a probability measure over X. As indicated, for a given LPS ρ I write ρ^j for the j-th probability measure in the sequence ρ. The **length** of ρ is denoted by $|\rho|$. I also write $\rho(j)$ for the **support** of ρ^j, that is $\rho(j) = \{x \in X : \rho^j(x) > 0\}$. An LPS ρ has **full support** iff $\cup\{\rho(i) : 1 \leq i \leq |\rho|\} = X$. Thus if ρ has full support, then every point x is in the support of some probability measure in ρ. Following [3, Definition 2], I write $x \geq_\rho x'$ if $\min\{j : x \in \rho(j)\} \leq \min\{j : x' \in \rho(j)\}$, and $x >_\rho x'$ if the inequality is strict. Informally, $x >_\rho x'$ means that the agent considers x more plausible than x', in that x is consistent with a "lower-order" belief than x'. To illustrate these definitions, Table 3 shows an LPS ρ_2 that might represent player 2's beliefs about 1's strategies in the Centipede Game, where $tt >_{\rho_2} lt >_{\rho_2} ll$. The LPS ρ_2 has full support.

Let S be a set of acts, and let $u : S \times X \to R$ be a utility function. The expected utility of an act s with respect to probability p and utility u is denoted by $EU(s, p, u)$ and defined as $EU(s, p, u) =_{df} \Sigma_{x \in X} p(x) \times u(s, x)$. The **lexicographic expected utility** of an act s with respect to an LPS ρ is a vector of real numbers (expected utilities) defined as $LEU(s, \rho, u) = (EU(s, \rho^1, u), EU(s, \rho^2, u), ..., EU(s, \rho^k, u))$ where $k = |\rho|$. For two vectors $\mathbf{u}, \mathbf{u}'$ of real numbers, let $\mathbf{u} \geq \mathbf{u}'$ denote the lexicographic ordering of the two vectors. Then an act s **maximizes lexicographic expected utility** given ρ, u iff $LEU(s, \rho, u) \geq LEU(s', \rho, u)$ for all $s' \in S$. A preference ordering $\succeq$ over S is **represented** by a pair (ρ, u) iff for all options s, s' we have that $s \succeq s' \iff LEU(s, \rho, u) \geq LEU(s', \rho, u)$. To illustrate, Table 4 shows the lexicographic expected utility of the strategies l_2 and t_2 for Player 2 in the Centipede game, given ρ_2. In this example, $LEU(l_2, \rho_2, u_2) = (0, 1, 2.5)$, and $EU(t_2, \rho_2, u_2) = (0, 2, 2)$, so if (ρ_2, u_2) represent the preferences of Player 2, then $t_2 \succ l_2$.

I next define the conditional LPS $\rho|P$ given a nonempty event P. As usual, if p is a probability measure over X, and $P \subseteq X$ an event such that $support(p) \cap P \neq \emptyset$, the conditional probability $p|P$ is defined by $[p|P](x) = p(x)/p(P)$ if $x \in P$, and $[p|P](x) = 0$ otherwise. Intuitively, to obtain the conditionalized probability system $\rho|P$, we first delete all probability measures ρ^j such that $support(\rho^j) \cap P = \emptyset$, and condition the remaining probability measures on P. For example, conditioning ρ_2 on $\{lt, ll\}$ yields $\rho_2|\{lt, ll\}$ displayed in Table 5. If $\rho|P$ does not have full support, the conditional probability

ρ_2^i/X	tt	tl	lt	ll	$EU(l_2, \rho_2^i, u_2)$	$EU(t_2, \rho_2^i, u_2)$
ρ_2^3	0	0	1/2	1/2	2.5	2
ρ_2^2	0	0	1	0	1	2
ρ_2^1	1/2	1/2	0	0	0	0

Table 4: The lexicographic expected utility of player 2's strategies in the Centipede game with utility function u_2 for player 2, given LPS ρ_2 from Table 3

$[\rho_2^i\|\{lt, ll\}]/X$	tt	tl	lt	ll
$[\rho_2^2\|\{lt, ll\}]$	0	0	1/2	1/2
$[\rho_2^1\|\{lt, ll\}]$	0	0	1	0

Table 5: The result of conditioning ρ_2 in Table 3 on $\{lt, ll\}$

$\rho|P$ may not be well-defined. According to decision theorists, an attractive feature of lexicographic probability systems with full support is that conditional probabilities are well-defined for any event [2]. The formal definition of $\rho|P$ is as follows.

1. $o(1) = \min\{1 \leq k \leq |\rho| : \rho(k) \cap S \neq \emptyset\}$. If ρ has full support, $o(1)$ is well-defined. And $(\rho|P)^1 = (\rho^{o(1)})|P$.

2. $o(n+1) = \min\{o(n) < k \leq |\rho| : \rho(k) \cap S \neq \emptyset\}$. If there is no such n, then $|\rho_P| = n$. Otherwise $(\rho|P)^{n+1} = (\rho^{o(n+1)})|P$.

I introduce a new operation $\rho * P = [\rho|P](1)$, which assigns to each event P the support of the first probability measure in $\rho|P$. I refer to this operation as the **revision** of ρ on P. For example, in the LPS ρ_2 above, we have that $\rho_2 * \{lt, ll\} = \{lt\}$ (see Table 5). A revision on P can be thought of as representing the "primary theory" or "first-order beliefs" of the agent given the information P.

Remark. As Stalnaker has noted [15, fn.12], lexicographic probability systems are closed related to structures that feature in the well-known AGM belief revision theory [5], [12]. Each LPS induces a revision operator + defined by $\rho(1) + P =_{df} \rho * P$; if we interpret this operation as a revision of the agent's "primary theory" $\rho(1)$ on the information P, it is easy to verify that the revision satisfies the well-known AGM axioms for minimal belief change. A difference is that in the AGM theory, a belief revision operator is a *binary* function from "current beliefs" K and "new information" P to new beliefs; in contrast, the revision associated with an LPS is a *unary* function of information P. (Hans Rott discusses advantages and disadvantages of unary vs. binary belief revision operators [10].)

4 General Epistemic Assumptions

Consider a 2-player game $G = \langle S_1, S_2, u_1, u_2 \rangle$, with sets of options S_1 and S_2, and utility functions u_1, u_2 defined on $S_1 \times S_2$. Let W be a set of states of the world. A given state of the world w associates the following elements with each player i:

1. a strategy choice $choice_i^w \in S_i$

2. a preference ordering $\succeq_i^w$ over the options S_i

3. a LPS ρ_i^w over S_{-i}, and hence a weak ordering $\geq_{\rho_i}^w$ over S_{-i}; I write more concisely $\succeq_i^w =_{df} \geq_{\rho_i}^w$.

4. a belief operator B_i^w. If A is an assertion about the game G, then $B_i^w(A)$ expresses the fact that in w, player i believes A.

One may take the belief operator B_i as given or interpret it in various ways, for example such that $B_i(A)$ represents probability 1 belief in A [13], or that $B_i(A)$ is the "first-order belief" of an agent in a lexicographic probability system, for example an LPS over a type space [1], or that $B_i(A)$ corresponds to "certain belief" [1] (see below). The theorems in this paper hold for any concept of belief that satisfies our axioms. In what follows, I consider the implications of various conditions on the epistemic elements listed above.

Definition 2 *Basic Epistemic Principles*

1. *(Lexicographic rationality)* ρ_i, u_i *represent* $\succeq_i$.
2. *(Full Support)* ρ_i *has full support.*
3. *(Preference Maximization) If* $choice_i = s_i$, *then* $\forall s_i'.s_i \succeq_i s_i'$.
4. *(Preference Introspection) If* $B_i(s_i \succeq s_i')$, *then* $s_i \succeq s_i'$.

I use the standard notion of **common belief** (see Section 1), which I denote by CB. Throughout this paper, I assume that common belief is closed under implication and that mathematical and logical truths are common belief. I also assume that *all aspects of the game*, or game tree, are common belief among the players.

5 Respect for Public Preferences and Proper Rationalizability

This section formalizes my key assumption, respect for public preferences. Let us consider again an agent A with three options, \$300, \$200, \$100. Suppose that an agent B believes that A's preference ranking is \$300 $\succ$ \$200 $\succ$ \$100. Then we may require that the first-order probability measure of B's lexicographic belief system conditional on the event $\{\$200, \$100\}$ assigns probability 1 to \$200. It is easy to see that this requirement is equivalent to the following axiom.

Axiom 3 *For each agent* i, *if* $B_i(s_{-i} \succ s_{-i}')$, *then* $s_{-i} >_i s_{-i}'$.

A weaker assumption is that this requirement holds only for preferences that are *public*, in the sense that they are common belief among the agents. My next axiom asserts that for public preferences, it is common belief that the preferred option is ranked above the dispreferred one.

Axiom 4 *(Respect for Public Preferences) For each agent* i, *if* $CB(s_i \succ_i s_i')$, *then* $CB(s_i >_{-i} s_i')$.

Given our assumptions about common belief, it is possible to derive Axiom 4 from common belief in Axiom 3.

Lemma 5 *Common belief in Axiom 3 implies Axiom 4.*

The remainder of this paper investigates the consequences of Respect for Public Preferences, given the general epistemic assumptions laid out in Section 4. Before I start this investigation, I clarify the relationship between my epistemic assumptions and previous work, particularly Schuhmacher and

Asheim's results on proper rationalizability. Reading the remainder of this section can be omitted without loss of continuity.

Axiom 3 is very closely related to Asheim's definition of "respecting preferences" [1, Sec. 4.1]. I outline Asheim's definition to clarify the precise relationship; for more details see [1]. The definition is given in a semantic framework with types. A state of the world is a tuple (s_1, s_2, t_1, t_2) where s_i is a pure strategy for player i and t_i is a type of player i from the set T_i of types of player i. In Asheim's framework, there are only finitely many types for each player [1, Def.1]. For each type $t_i \in T_i$ there is a preference relation $\succeq^{t_i}$over the pure strategies of player i, and an LPS ρ^{t_i} over the points in $S_{-i} \times T_{-i}$. Define the events $[t_i] = \{(s_i, t_i) : s_i \in S_i\}$ and $[s_i] = \{(s_i, t_i) : t_i \in T_i\}$. For a given lexicographic system ρ over points X, define *certain belief* B_ρ by $B_\rho(E)$ iff $supp(\rho) \cap E = \emptyset$. In other words, E is certain belief given ρ iff E receives probability 0 in every measure in ρ. The dual *possibility operator* P_ρ associated with an LPS ρ is given by $P_\rho(E) \iff \neg B_\rho(\overline{E})$. Asheim introduces a *cautiousness* condition for players' beliefs. The event that player i is cautious is defined by: $(s_1, s_2, z_1, z_2) \in cau_i \iff$ for all types $t_{-i} \in T_{-i}$, if $P_{\rho^{z_i}}([t_{-i}])$, then $P_{\rho^{z_i}}(\{(s_{-i}, t_{-i})\})$ for all $s_{-i} \in S_{-i}$. So player i is cautious if for each type t_{-i} that i considers epistemically possible, player i considers all strategies $s_{-i} \in S_{-i}$ (i.e., all pairs (s_{-i}, t_{-i})) epistemically possible.

In this notation, Asheim's definition of the event that player i respects preferences corresponds to: $(s_1, s_2, z_1, z_2) \in resp_i \iff$ for all pairs $(s_{-i}, t_{-i}), (s'_{-i}, t_{-i}) \in S_{-i} \times T_{-i}$, if $P_{\rho^{z_i}}([t_{-i}])$ and $s_{-i} \succ^{t_{-i}} s'_{-i}$, then $(s_{-i}, t_{-i}) >_{\rho^{z_i}} (s'_{-i}, t_{-i})$. To illustrate, suppose that player 1 considers epistemically possible a type t_2 of player 2 (i.e., $P_{\rho^{t_1}}([t_2])$), and that type t_2 prefers strategy s_2 to s'_2. Then player 1 ranks the pair (s_2, t_2) higher than (s'_2, t_2) in his LPS ρ^{t_1}. The intuition is very close to that behind Axiom 3: given that player 2's prefers s_2 to s'_2 (i.e., given that his type is t_2), player 1 ranks s_2 higher than s'_2. To see how our results relate to Asheim's axioms, interpret the belief operator B_i as certain belief $B_{\rho^{t_i}}$, and the lexicographic probability system ρ_i as the marginal $\rho^{t_i}_m$ of ρ^{t_i}, which is defined by $(\rho^{t_i}_m)^j(s_{-i}) =_{df} \Sigma_{t_{-i} \in T_{-i}}(\rho^{t_i})^j(s_{-i})$. With this interpretation, *cautiousness and respect for preferences imply Axiom 3.* (I omit the straightforward proof.)

Respect for public preferences (Axiom 4) is weaker than respect for preferences because it applies only to preferences that are common belief among the agents. Thus even if, for example, player 1 has certain belief that player 2 prefers s_2 to s'_2, but this fact is not common (certain) belief, then Axiom 4 does not require player 1 to respect the preference of s_2 over s'_2 (i.e., Axiom 4 allows that $s'_2 >_{\rho^{t_1}} s_2$). As it turns out, the weak condition of respect for public preferences is sufficient to validate Iterated Backward Inference, the elimination procedure presented in this paper.

Asheim refers to the conjunction of cautiousness and respect for preferences (plus knowledge of the game structure) as *proper consistency* [1, Sec. 4.1]. A strategy s_i is *properly rationalizable* if s_i maximizes preferences in a state of the world in which there is common (certain) belief of proper consistency [1, Def. 2]. Asheim proves that this definition of proper rationalizability is equivalent to the previous one by Schuhmacher [1, Prop. 3], [11]. Since common belief in proper consistency entails common belief in Axiom 3, which in turn entails respect for public preferences, the upshot is that Iterated Backward Inference can be used to compute properly rationalizable strategies: if the procedure eliminates a strategy s_i, then s_i is not properly rationalizable.

6 An Iterated Elimination Procedure for Respect for Public Preferences

This section defines the algorithm for deriving consequences of respect for public preferences.

6.1 Sequential Admissibility and Entailment Inference

The iterated elimination procedure for deriving consequences of Respect for Public Preferences combines two principles: "local" dominance at an information set, and backward inference. I begin with dominance at an information set. To simplify definitions, I focus on games with *perfect recall.* Intuitively, a game has perfect recall if no player forgets what they once did or knew. A formal definition may be found in any text on game theory, for example in [9, p.203].

Definition 6 *Let T be a game tree with perfect recall and information set I_i belonging to player i.*

1. *s_i weakly dominates s_i' at I_i iff*

   (a) *s_i and s_i' are each consistent with I_i (i.e., $s_i, s_i' \in [I_i]_i$), and*

   (b) *there is a strategy $s_{-i} \in S_{-i}(T)$ consistent with I_i such that $u_i(s_i, s_{-i}) > u_i(s_i', s_{-i})$, and*

   (c) *$u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i})$ for all s_{-i} consistent with I_i.*

2. *s_i strictly dominates s_i' at I_i given $\Sigma_{-i} \subseteq S_{-i}(T)$ iff*

   (a) *s_i and s_i' are each consistent with I_i, and*

   (b) *I_i is consistent with Σ_{-i} (i.e., $[I_i]_{-i} \cap \Sigma_{-i} \neq \emptyset$), and*

   (c) *$u_i(s_i, s_{-i}) > u_i(s_i', s_{-i})$ for all $s_{-i} \in [I_i]_{-i} \cap \Sigma_{-i}$.*

To illustrate, in the Centipede game (Figure 1) for player 1 the strategy lt weakly dominates ll at information set I^1, and strictly dominates ll at I^3 given $\{l_2, t_2\}$. The strategy tt strictly dominates both lt and ll at I^1 given $\{t_2\}$. For player 2, the strategy t_2 strictly dominates l_2 at I^2 given $\{lt\}$. In Kohlberg's game (Figure 2), for player II the strategy UL strictly dominates DL at information set I^1 given $S_I(T)$. At I^2, the strategy yT strictly dominates nB given $S_{II}(T)$, and yT strictly dominates nT given $\{DR\}$. Finally, at I^4 the strategy nB strictly dominates nT given $\{DR\}$.

The second component of our iterated elimination procedure draws inferences from the results of elimination at one information set I to eliminate strategies at another information set I'. Consider two information sets I, I' such that all strategies s_i for player i consistent with I are also consistent with I'; in symbols $[I]_i \subseteq [I']_i$. In this case I say that I **entails** I' **for player** i. The general principle is this: if a strategy s_i consistent with I is considered unlikely given the information in the set I, and I entails I' for player i, then s_i is considered unlikely given the information in the set I'. Intuitively, if s_i is considered unlikely given I, then there is a possibility s_i' consistent with I that is considered more likely than s_i. Since I entails I', the possibility s_i' is consistent with I' and hence s_i is not among the most likely possibilities given I'. To give the principle a precise formulation, I interpret "s_i is considered unlikely at information set I" to mean that the revision on I rules out the strategy s_i. Then in symbols, the **entailment inference principle** is that

$$\text{if } [I]_i \subseteq [I']_i, \text{and } s_i \in [I]_i, \text{but } s_i \notin \rho_{-i} * [I]_i, \text{then } s_i \notin \rho_{-i} * [I']_i. \tag{1}$$

Or contrapositively: If I entails I' for player i, then $(\rho_{-i} * [I']_i) \cap [I]_i \subseteq \rho_{-i} * [I]_i$. For example, in the Centipede game, if $lt \notin \rho_2 * [I^3]_1$, then the entailment inference principle implies that $lt \notin \rho_2 * [I^2]_1$. In Kohlberg's game, if $nT \notin \rho_I * [I^4]_{II}$, then the principle implies that $nT \notin \rho_I * [I^2]_{II}$. On the other hand, in Kohlberg's game it is not the case that $[I^1]_2 \subseteq [I^2]_2$, so even if DL is considered unlikely at I^1, the principle does not allow us to infer that DL is considered unlikely at I^2. For example, we may have that $\rho_1 * [I^1]_2 = \{UL, UR\}$ and $\rho_1 * [I^2]_2 = \{DL, DR\}$. Thus Principle 1 does not in general license "forward induction" arguments.

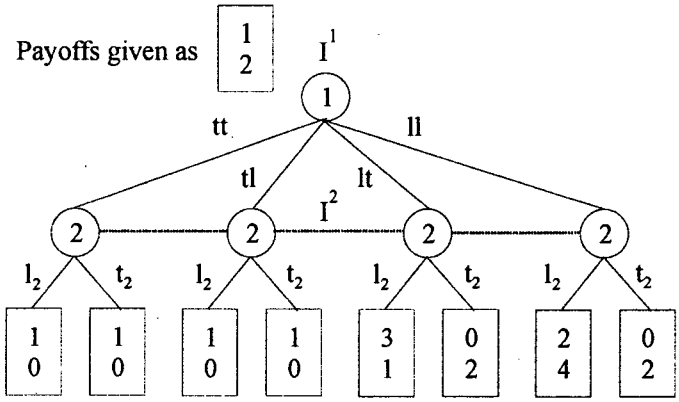

Figure 3: A game tree corresponding to the strategic form of the Centipede game

Lexicographic revision satisfies Principle 1. The next Lemma shows that this is due to a very general property of lexicographic probability systems.

Lemma 7 *Let ρ be an LPS with full support, and suppose that $P \subseteq P'$. Then $(\rho * P') \cap P \subseteq \rho * P$.*

If we set $[I]_i = P$, and $[I']_i = P'$, it is apparent that Principle 1 is an instance of the contrapositive of Lemma 7. Standard backward inference can be seen as an instance of Principle 1. Intuitively, in backward inference a player "looks ahead" and "reasons back". Consider an information set I_i belonging to player i that follows some other information set I, which may belong to player $-i$ (in symbols, $I_i \geq I$). It is possible to show that for games with perfect recall, in this case I_i entails I for player i. So by Principle 1, we have that if a strategy s_i consistent with I is considered unlikely given the information in the set I_i, then it is considered unlikely given the information in the set I. In symbols, in *games with perfect recall* Principle 1 implies that

$$\text{if } I_i > I, \text{ and } s_i \in [I]_i, \text{but } s_i \notin \rho_{-i} * [I_i]_i, \text{then } s_i \notin \rho_{-i} * [I]_i. \tag{2}$$

I refer to Principle 2 as the **backward inference principle**. Backward inference is the typical application of Principle 1. A different case in which Principle 1 applies occurs when we have two information sets I, I' such that $I' > I$, and *all* strategies consistent with information set I are consistent with I'—that is, $[I] \subseteq [I']$, and hence $[I]_i \subseteq [I']_i$. This special case arises in a game with just two information sets, such as results from transcribing a game matrix directly into a game tree. For example, Figure 3 shows a game tree in which players' options are just their strategies in the Centipede game. If for example $l_2 \notin \rho_1 * [I^1]_2$, then $l_2 \notin \rho_1 * [I^2]_2$; in general, we have that $\rho_1 * [I^2]_2 \subseteq \rho_1 * [I^1]_2$.

6.2 Iterated Backward Inference: Definition and Examples

Now I define the iterated elimination procedure, which I refer to as Iterated Backward Inference, or IBI for short.

Definition 8 (Iterated Backward Inference) *Let T be a game tree with perfect recall. Then define for all players i, for all information sets I_i, strategies $s_i \in S_i(T), s_{-i} \in S_{-i}(T)$:*

1. $s_i \in \Gamma_i^0(I) \iff$

information set/ surviving strategies	I^1	I^2	I^3
Γ_1^0	tt, tl, lt	lt	lt
Γ_2^0	l_2, t_2	l_2, t_2	l_2
Γ_1^1	tt, tl, lt	lt	lt
Γ_2^1	t_2	t_2	l_2
Γ_1^2	tt, tl	lt	lt
Γ_2^2	t_2	t_2	l_2

Table 6: Iterated Backward Inference in the Centipede Game (see Figure 1)

information set/ surviving strategies	I^1	I^2
Γ_1^0	tt, tl, lt	tt, tl, lt
Γ_2^0	l_2, t_2	l_2, t_2

Table 7: Iterated Backward Inference in a matrix game tree for the Centipede game (see Figure 3)

(a) $s_i \in [I]_i$, *and*

(b) *for all information sets* I_i *belonging to player* i *such that* $[I_i]_i \subseteq [I]_i$, *we have that* s_i *is not weakly dominated at* I_i.

2. $s_i \in \Gamma_i^{n+1}(I) \iff$

(a) $s_i \in \Gamma_i^n(I)$, *and*

(b) *for all information sets* I_i *belonging to player* i *such that* $[I_i]_i \subseteq [I]_i$, *we have that* s_i *is not strictly dominated at* I_i *given* $\Gamma_{-i}^n(I_i)$.

For a game tree T with root r and information set I^r containing r, let $\Gamma_i^n(T) =_{df} \Gamma_i^n(I_r)$. In a finite game tree, it is clear that there is some round m after which no more elimination takes place, i.e., $\Gamma_i^m(T) = \Gamma_i^{m+1}(T)$ for all i; I write $\Gamma_i^\infty(T) =_{df} \Gamma_i^m(T)$. To illustrate IBI, I show its computations on the games of Figures 1, 3 and 2. The columns correspond to information sets in the game and the rows show the set of strategies $\Gamma_i^n, \Gamma_{-i}^n$ surviving at each round of elimination, for each player. Table 6 shows that the final result of the computation in the Centipede game is for Player 1 to take at all his information sets, and for player 2 to take at his. Formally, we have that $\Gamma_1^\infty(T) = \{tt, tl\}$ and $\Gamma_2^\infty(T) = \{t_2\}$.Table 7 shows that Iterated Backward Inference eliminates just one strategy in the matrix tree (normal form) of the Centipede game, namely ll which is weakly dominated by lt. After ll is dominated, there is no strict dominance, and the procedure terminates. This example illustrates two points. First, in games with two information sets, which are essentially just another representation of the strategic form of a game, Iterated Backward Inference coincides with the Dekel-Fudenberg procedure: First eliminate all weakly dominated strategies, then iteratively eliminate strictly dominated strategies. In light of Theorem 9 below, it follows that the Dekel-Fudenberg procedure is valid for deriving consequences of Respect for Public Preferences. The second point is that Iterated Backward Inference can yield different results for game trees that have the same normal form, as the two game trees for the Centipede Game do. In each case, the output of Iterated Backward Inference is valid in the sense that Respect for Public Preferences entails that eliminated strategies will not be played. In our examples, applying IBI in the game tree of Figure 3 yields the result that Player 1 does not choose ll, and applying IBI in the game tree of Figure

information set/ surviving strategies	I^1	I^2	I^3	I^4
Γ_1^0	yT, yB, nT	yT, yB, nT	nT, nB	nT, nB
Γ_2^0	UL, UR, DR	DL, DR	DL, DR	DL, DR
Γ_1^1	yT, yB, nT	yT, yB, nT	nT, nB	nT, nB
Γ_2^1	UL, UR	DL, DR	DL, DR	DL, DR

Table 8: Iterated Backward Inference in Kohlberg's game (Figure 2)

1 yields the result that Player 1 chooses neither ll nor lt. Thus in some game trees IBI provides more information than in others, in that the procedure finds more of the consequences of Respect for Public Preferences. So although our main *epistemic principle*, Respect for Public Preferences, pertains to the strategic form of a game, and hence is independent of any particular extensive form, the *computational procedure* of this paper does depend on a particular choice of game tree.

Table 8 shows the computation of Iterated Backward Inference in Kohlberg's game. This example illustrates two points. First, even though Kohlberg's game is not a game of perfect information, Iterated Backward Inference yields a unique outcome prediction: that player II will choose U immediately, resulting in payoffs $(0, 2)$. Second, the game shows how Iterated Backward Inference incorporates backward reasoning but not forward reasoning. At information set I^2, both yT and yB strictly dominate nB, so nB is eliminated at I^2 at round 0, and hence by backward inference, nB is also eliminated at I^1 at round 0. After nB is eliminated, UL and UR strictly dominate DR at information set I^1 in round 1, leaving U as the only choice for player II. By contrast, there are two forward induction arguments that IBI does not incorporate. First, since DL is eliminated at I^1, one might take forward induction to imply that DL should be eliminated at I^2 as well. Second, since nB is eliminated at I^2, one might take forward induction to imply that nB should be eliminated at I^3 as well.

7 Soundness, Existence and Backward Induction

This section contains the main theorems of the paper. First, the key result: For finite games with perfect recall, if respect for public preferences obtains and lexicographic rationality with full support is common belief, then it is common belief that each player believes that play follows Iterated Backward Inference, at each information set I. Secondly existence: in finite games there are some predictions consistent with IBI. Third, IBI generalizes backward induction in perfect information games.

Theorem 9 *Let T be a finite game tree with perfect recall and assume that Lexicographic Rationality and Full Support are common belief (see Axiom 2). Then Respect for Public Preferences (Axiom 4) implies that for all n, i, I:*

1. *if a strategy s_i is strictly dominated at an information set I given $\Gamma^n_{-i}(I)$, then $CB(\exists s'_i \in [I]_i . s'_i \succ_i s_i)$, and*
2. *$CB(\rho_{-i} * [I]_i \subseteq \Gamma^n_i(I))$.*

Recall that $choice_i$ denotes the strategy choice of player i. It is a simple corollary from Theorem 9 that IBI makes correct predictions about the choices of the players, given common belief in revealed preference and our standard epistemic assumptions.

Corollary 10 *Let T be a finite game tree with perfect recall and assume that Lexicographic Rationality and Full Support are common belief. Then Respect for Public Preferences (Axiom 4), Preference Maximization and Introspection (see Definition 2) imply that $choice_i \in \Gamma_i^n(T)$ for all n.*

In finite games with perfect recall, IBI returns a nonempty result.

Proposition 11 *Let T be a finite game tree with perfect recall. Then $\Gamma^n(T) \neq \emptyset$ for all n.*

In finite perfect information games or repeated stage games in which backward induction yields a unique solution, IBI agrees with the backward induction solution. Thus we have the following result.

Proposition 12 *Let T be a finite game tree with perfect information that has a unique subgame perfect equilibrium (s_i, s_{-i}). Then $u_i(s_i, s_{-i}) = u_i(s'_i, s'_{-i})$ for every strategy profile $(s'_i, s'_{-i}) \in \Gamma^\infty(T)$ and each player i.*

The definition of subgame perfect equilibrium may be found in any text on game theory, for example in [9, Ch.6.2]. For further discussion of the relationship between backward induction, respect for preferences and proper rationalizability see [11] and [1].

8 Conclusion

An important approach to developing and understanding solution concepts for game theory is to examine the epistemic assumptions that underlie predictions about the outcome of a game. In this paper I considered the consequences of Respect for Public Preferences: if it is common belief that an agent A prefers option a to option b, then it is common belief that, given that A chooses either a or b, she chooses a. Following previous work by [3] and [1], I proposed to apply lexicographic probability systems and capture Respect for Public Preferences by requiring that each agent $B \neq A$ assigns probability 1 to A's choosing a, conditional on A choosing a or b, whenever A's preference for a over b is common belief.

Iterated Backward Inference (IBI) is a procedure for computing the consequences of Respect for Public Preferences in a given game G. IBI eliminates strategies in a game tree T. The main result is that the procedure is valid given common belief in Revealed Preference, in the following sense: if T is an extensive form of the strategic game G, and IBI eliminates a strategy s, then s is not chosen in the game G. [IBI generalizes two well-known algorithms for solving games: the Dekel-Fudenberg procedure (first eliminate weakly dominated strategies, then iteratively strictly dominates ones), and standard backward induction for game trees with perfect information; IBI yields predictions that are at least as strong as those given by these two algorithms.] It follows from Asheim's characterization of proper rationalizability [1] that properly rationalizable strategies are consistent with Respect for Public Preferences. Hence IBI can be used to find strategies that are not properly rationalizable.

The fact that IBI is valid for computing consequences of common belief in Revealed Preference rests on two key facts. First, given perfect recall, lexicographic rationality enforces sequential admissibility (admissibility at reachable information sets). Second, lexicographic rationality satisfies the entailment inference principle: Consider two information sets I, I' such that all strategies consistent with I are consistent with I'. Then if a strategy s is considered unlikely at I, the strategy s is also considered unlikely at I'. When I follows I' in a game with perfect recall, the entailment inference principle yields a backward inference principle.

I mention two open questions for future research. In different game trees with the same strategic form G, Iterated Backward Inference may give different (stronger) results. We would like to apply IBI to a canonical game tree $T(G)$ in which the procedures gives complete results, eliminating all and only those

strategies inconsistent with Respect for Public Preferences. The examples in this paper suggest that canonical game trees are those that in some sense have as many information sets as possible. Whether canonical game trees exist for an arbitrary game and how to construct them is the most important open question for understanding the computational aspects of Respect for Public Preferences, and perhaps of proper rationalizability as well.

Respect for Public Preferences does not validate typical forward induction arguments (for example, it does not entail the forward induction solution in the well-known Burning Dollar game). One would like to know further epistemic principles that underlie forward induction arguments. Is there a "best rationalization" principle based on Respect for Public Preferences that validates forward induction?

9 Acknowledgements

I am indebted to Geir Asheim, Mamoru Kaneko, Cristina Bicchieri and Phil Curry for helpful discussions. Anonymous referees for TARK provided useful comments. Research on this paper was supported by a grant from the Social Sciences and Humanities Research Council of Canada.

References

[1] Geir Asheim. Proper rationalizability in lexicographic beliefs. *International Journal of Game Theory*, 30:453:478, 2001.

[2] Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59(1):61–79, 1991.

[3] Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and equilibrium refinements. *Econometrica*, 59(1):81–98, 1991.

[4] Eddie Dekel and Drew Fudenberg. Rational behavior with payoff uncertainty. *Journal of Economic Theory*, 52:243–267, 1990.

[5] Peter Gärdenfors. *Knowledge In Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, Mass., 1988.

[6] Mamoru Kaneko and J. Jude Kline. Behavior strategies, mixed strategies and perfect recall. *International Journal of Game Theory*, 24:127–145, 1995.

[7] Elon Kohlberg. Refinement of nash equilibrium: The main ideas. In *Game Theory and Applications.* Academic Press, San Diego, 1990.

[8] R. Myerson. Refinements of the nash equilibrium concept. *International Journal of Game Theory*, 7:73–80, 1978.

[9] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory.* MIT Press, Cambridge, Mass., 1994.

[10] Hans Rott. Coherence and conservatism in the dynamics of belief. part i: Finding the right framework. *Erkenntnis*, 50:387–412, 1999.

[11] F. Schuhmacher. Proper rationalizability and backward induction. *International Journal of Game Theory*, 28:599–615, 1999.

[12] Oliver Schulte. Minimal belief change, pareto-optimality and logical consequence. *Economic Theory*, 19(1):105–144, 2002.

[13] Robert Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[14] Robert Stalnaker. On the evaluation of solution concepts. *Theory and Decision*, 37:49–73, 1996.

[15] Robert Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.