Logical agents

Chapter 7
Outline

♦ Knowledge-based agents
♦ Wumpus world
♦ Logic in general—models and entailment
♦ Propositional (Boolean) logic
♦ Equivalence, validity, satisfiability
♦ Inference rules and theorem proving
  - forward chaining
  - backward chaining
  - resolution
Knowledge bases

- Inference engine
- Knowledge base

- domain-independent algorithms
- domain-specific content

Knowledge base = set of sentences in a formal language

**Declarative** approach to building an agent (or other system):

TELL it what it needs to know

Then it can ASK itself what to do—answers should follow from the KB

Agents can be viewed at the **knowledge level**

i.e., what they know, regardless of how implemented

Or at the **implementation level**

i.e., data structures in KB and algorithms that manipulate them
A simple knowledge-based agent

function KB-AGENT(\textit{percept}) returns an \textit{action} 
  static: $KB$, a knowledge base 
            $t$, a counter, initially 0, indicating time 

  \textsc{Tell}(KB, \textsc{Make-Percept-Sentence}(\textit{percept}, t)) 
  \textit{action} $\leftarrow$ \textsc{Ask}(KB, \textsc{Make-Action-Query}(t)) 
  \textsc{Tell}(KB, \textsc{Make-Action-Sentence}(\textit{action}, t)) 
  $t \leftarrow t + 1$

return \textit{action}

The agent must be able to:
  Represent states, actions, etc.
  Incorporate new percepts
  Update internal representations of the world
  Deduce hidden properties of the world
  Deduce appropriate actions
Wumpus World PEAS description

Performance measure
  gold +1000, death -1000
  -1 per step, -10 for using the arrow

Environment
  Squares adjacent to wumpus are smelly
  Squares adjacent to pit are breezy
  Glitter iff gold is in the same square
  Shooting kills wumpus if you are facing it
  Shooting uses up the only arrow
  Grabbing picks up gold if in same square
  Releasing drops the gold in same square

Sensors Breeze, Glitter, Smell

Actuators Left turn, Right turn,
  Forward, Grab, Release, Shoot
Wumpus world characterization

Observable??
Observable? No—only local perception

Deterministic??
Wumpus world characterization

**Observable??** No—only *local* perception

**Deterministic??** Yes—outcomes exactly specified

**Episodic??**
Wumpus world characterization

Observable?? No—only local perception
Deterministic?? Yes—outcomes exactly specified
Episodic?? No—sequential at the level of actions
Static??
<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable</td>
<td>No—only local perception</td>
</tr>
<tr>
<td>Deterministic</td>
<td>Yes—outcomes exactly specified</td>
</tr>
<tr>
<td>Episodic</td>
<td>No—sequential at the level of actions</td>
</tr>
<tr>
<td>Static</td>
<td>Yes—Wumpus and Pits do not move</td>
</tr>
<tr>
<td>Discrete</td>
<td></td>
</tr>
</tbody>
</table>
**Wumpus world characterization**

**Observable??**  No—only local perception

**Deterministic??**  Yes—outcomes exactly specified

**Episodic??**  No—sequential at the level of actions

**Static??**  Yes—Wumpus and Pits do not move

**Discrete??**  Yes

**Single-agent??**
Wumpus world characterization

Observable?? No—only local perception

Deterministic?? Yes—outcomes exactly specified

Episodic?? No—sequential at the level of actions

Static?? Yes—Wumpus and Pits do not move

Discrete?? Yes

Single-agent?? Yes—Wumpus is essentially a natural feature
Exploring a wumpus world

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OK</td>
<td>A</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chapter 7  13
Exploring a wumpus world

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>OK</td>
<td></td>
<td>OK</td>
<td>OK</td>
<td></td>
</tr>
</tbody>
</table>

Chapter 7  14
Exploring a wumpus world
Exploring a wumpus world
Exploring a wumpus world
Exploring a wumpus world

P

B OK

A

A OK S OK

A W
Exploring a wumpus world
Exploring a wumpus world
Other tight spots

Breeze in (1,2) and (2,1)
⇒ no safe actions

Assuming pits uniformly distributed,
(2,2) has pit w/ prob 0.86, vs. 0.31

Smell in (1,1)
⇒ cannot move

Can use a strategy of coercion:
shoot straight ahead
wumpus was there ⇒ dead ⇒ safe
wumpus wasn’t there ⇒ safe
Logic in general

Logics are formal languages for representing information such that conclusions can be drawn.

Syntax defines the sentences in the language.

Semantics define the “meaning” of sentences; i.e., define truth of a sentence in a world.

E.g., the language of arithmetic:

$x + 2 \geq y$ is a sentence; $x^2 + y >$ is not a sentence.

$x + 2 \geq y$ is true iff the number $x + 2$ is no less than the number $y$.

$x + 2 \geq y$ is true in a world where $x = 7, y = 1$.

$x + 2 \geq y$ is false in a world where $x = 0, y = 6$. 

**Entailment**

Entailment means that one thing follows from another:

\[ KB \models \alpha \]

Knowledge base \( KB \) entails sentence \( \alpha \)

if and only if

\( \alpha \) is true in all worlds where \( KB \) is true

E.g., the KB containing “the Giants won” and “the Reds won” entails “Either the Giants won or the Reds won”

E.g., \( x + y = 4 \) entails \( 4 = x + y \)

Entailment is a relationship between sentences (i.e., syntax) that is based on semantics

Note: brains process syntax (of some sort)
Models

Logicians typically think in terms of models, which are formally structured worlds with respect to which truth can be evaluated.

We say $m$ is a model of a sentence $\alpha$ if $\alpha$ is true in $m$.

$M(\alpha)$ is the set of all models of $\alpha$.

Then $KB \models \alpha$ if and only if $M(KB) \subseteq M(\alpha)$.

E.g. $KB = \text{Giants won and Reds won}$

$\alpha = \text{Giants won}$
Entailment in the wumpus world

Situation after detecting nothing in [1,1], moving right, breeze in [2,1]

Consider possible models for ?s assuming only pits

3 Boolean choices $\Rightarrow$ 8 possible models
Wumpus models
Wumpus models

\[ KB = \text{wumpus-world rules} + \text{observations} \]
\[ \text{KB} = \text{wumpus-world rules} + \text{observations} \]

\[ \alpha_1 = \text{“[1,2] is safe”, KB } \models \alpha_1, \text{ proved by model checking} \]
$KB = \text{wumpus-world rules} + \text{observations}$
\[ K B = \text{wumpus-world rules} + \text{observations} \]

\[ \alpha_2 = \text{“[2,2] is safe”, } K B \not\models \alpha_2 \]
Inference

\[ KB \vdash_i \alpha \] = sentence \( \alpha \) can be derived from \( KB \) by procedure \( i \)

Consequences of \( KB \) are a haystack; \( \alpha \) is a needle.
Entailment = needle in haystack; inference = finding it

**Soundness:** \( i \) is sound if
\[
\text{whenever } KB \vdash_i \alpha, \text{ it is also true that } KB \models \alpha
\]

**Completeness:** \( i \) is complete if
\[
\text{whenever } KB \models \alpha, \text{ it is also true that } KB \vdash_i \alpha
\]

Preview: we will define a logic (first-order logic) which is expressive enough to say almost anything of interest, and for which there exists a sound and complete inference procedure.

That is, the procedure will answer any question whose answer follows from what is known by the \( KB \).
Propositional logic: Syntax

Propositional logic is the simplest logic—illuminates basic ideas.

The proposition symbols \( P_1, P_2 \) etc are sentences.

If \( S \) is a sentence, \( \neg S \) is a sentence \((\text{negation})\)

If \( S_1 \) and \( S_2 \) are sentences, \( S_1 \land S_2 \) is a sentence \((\text{conjunction})\)

If \( S_1 \) and \( S_2 \) are sentences, \( S_1 \lor S_2 \) is a sentence \((\text{disjunction})\)

If \( S_1 \) and \( S_2 \) are sentences, \( S_1 \implies S_2 \) is a sentence \((\text{implication})\)

If \( S_1 \) and \( S_2 \) are sentences, \( S_1 \iff S_2 \) is a sentence \((\text{biconditional})\)
Propositional logic: Semantics

Each model specifies true/false for each proposition symbol

E.g. \( P_{1,2} \quad P_{2,2} \quad P_{3,1} \)  
\( \text{true} \quad \text{true} \quad \text{false} \)

(With these symbols, 8 possible models, can be enumerated automatically.)

Rules for evaluating truth with respect to a model \( m \):

\[ \neg S \text{ is true iff } \quad S \text{ is false} \]
\[ S_1 \land S_2 \text{ is true iff } \quad S_1 \text{ is true and } \quad S_2 \text{ is true} \]
\[ S_1 \lor S_2 \text{ is true iff } \quad S_1 \text{ is true or } \quad S_2 \text{ is true} \]
\[ S_1 \Rightarrow S_2 \text{ is true iff } \quad S_1 \text{ is false or } \quad S_2 \text{ is true} \]
\[ \text{i.e., is false iff } \quad S_1 \text{ is true and } \quad S_2 \text{ is false} \]
\[ S_1 \Leftrightarrow S_2 \text{ is true iff } \quad S_1 \Rightarrow S_2 \text{ is true and } \quad S_2 \Rightarrow S_1 \text{ is true} \]

Simple recursive process evaluates an arbitrary sentence, e.g.,
\[ \neg P_{1,2} \land (P_{2,2} \lor P_{3,1}) = \text{true} \land (\text{false} \lor \text{true}) = \text{true} \land \text{true} = \text{true} \]
## Truth tables for connectives

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$\neg P$</th>
<th>$P \land Q$</th>
<th>$P \lor Q$</th>
<th>$P \Rightarrow Q$</th>
<th>$P \Leftrightarrow Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
</tbody>
</table>
Wumpus world sentences

Let $P_{i,j}$ be true if there is a pit in $[i, j]$.
Let $B_{i,j}$ be true if there is a breeze in $[i, j]$.

$\neg P_{1,1}$
$\neg B_{1,1}$
$B_{2,1}$

“Pits cause breezes in adjacent squares”
Wumpus world sentences

Let \( P_{i,j} \) be true if there is a pit in \([i,j]\).
Let \( B_{i,j} \) be true if there is a breeze in \([i,j]\).

\[-P_{1,1}\]
\[-B_{1,1}\]
\[B_{2,1}\]

“Pits cause breezes in adjacent squares”

\[B_{1,1} \iff (P_{1,2} \lor P_{2,1})\]
\[B_{2,1} \iff (P_{1,1} \lor P_{2,2} \lor P_{3,1})\]

“A square is breezy if and only if there is an adjacent pit”
### Truth tables for inference

<table>
<thead>
<tr>
<th>$B_{1,1}$</th>
<th>$B_{2,1}$</th>
<th>$P_{1,1}$</th>
<th>$P_{1,2}$</th>
<th>$P_{2,1}$</th>
<th>$P_{2,2}$</th>
<th>$P_{3,1}$</th>
<th>$KB$</th>
<th>$\alpha_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>
Inference by enumeration

Depth-first enumeration of all models is sound and complete

```
function TT-ENTAILS?(KB, α) returns true or false
    symbols ← a list of the proposition symbols in KB and α
    return TT-CHECK-ALL(KB, α, symbols, [])

function TT-CHECK-ALL(KB, α, symbols, model) returns true or false
    if EMPTY?(symbols) then
        if PL-TRUE?(KB, model) then return PL-TRUE?(α, model)
        else return true
    else do
        P ← FIRST(symbols); rest ← REST(symbols)
        return TT-CHECK-ALL(KB, α, rest, EXTEND(P, true, model)) and
        TT-CHECK-ALL(KB, α, rest, EXTEND(P, false, model))
```

$O(2^n)$ for $n$ symbols; problem is co-NP-complete
Logical equivalence

Two sentences are logically equivalent iff true in same models:
\[ \alpha \equiv \beta \quad \text{if and only if} \quad \alpha \models \beta \text{ and } \beta \models \alpha \]

\[
\begin{align*}
(\alpha \land \beta) & \equiv (\beta \land \alpha) \quad \text{commutativity of } \land \\
(\alpha \lor \beta) & \equiv (\beta \lor \alpha) \quad \text{commutativity of } \lor \\
((\alpha \land \beta) \land \gamma) & \equiv (\alpha \land (\beta \land \gamma)) \quad \text{associativity of } \land \\
((\alpha \lor \beta) \lor \gamma) & \equiv (\alpha \lor (\beta \lor \gamma)) \quad \text{associativity of } \lor \\
\neg(\neg\alpha) & \equiv \alpha \quad \text{double-negation elimination} \\
(\alpha \Rightarrow \beta) & \equiv (\neg \beta \Rightarrow \neg \alpha) \quad \text{contraposition} \\
(\alpha \Rightarrow \beta) & \equiv (\neg \alpha \lor \beta) \quad \text{implication elimination} \\
(\alpha \Leftrightarrow \beta) & \equiv ((\alpha \Rightarrow \beta) \land (\beta \Rightarrow \alpha)) \quad \text{biconditional elimination} \\
\neg(\alpha \land \beta) & \equiv (\neg \alpha \lor \neg \beta) \quad \text{de Morgan} \\
\neg(\alpha \lor \beta) & \equiv (\neg \alpha \land \neg \beta) \quad \text{de Morgan} \\
(\alpha \land (\beta \lor \gamma)) & \equiv ((\alpha \land \beta) \lor (\alpha \land \gamma)) \quad \text{distributivity of } \land \text{ over } \lor \\
(\alpha \lor (\beta \land \gamma)) & \equiv ((\alpha \lor \beta) \land (\alpha \lor \gamma)) \quad \text{distributivity of } \lor \text{ over } \land 
\end{align*}
\]
Validity and satisfiability

A sentence is valid if it is true in all models,
e.g., True, \(A \lor \neg A\), \(A \Rightarrow A\), \((A \land (A \Rightarrow B)) \Rightarrow B\)

Validity is connected to inference via the Deduction Theorem:
\(KB \models \alpha\) if and only if \((KB \Rightarrow \alpha)\) is valid

A sentence is satisfiable if it is true in some model
e.g., \(A \lor B\), \(C\)

A sentence is unsatisfiable if it is true in no models
e.g., \(A \land \neg A\)

Satisfiability is connected to inference via the following:
\(KB \models \alpha\) if and only if \((KB \land \neg \alpha)\) is unsatisfiable
Proof methods

Proof methods divide into (roughly) two kinds:

Application of inference rules
- Legitimate (sound) generation of new sentences from old
- **Proof** = a sequence of inference rule applications
  Can use inference rules as operators in a standard search alg.
- Typically require translation of sentences into a normal form

Model checking
 truth table enumeration (always exponential in $n$)
 improved backtracking, e.g., Davis–Putnam–Logemann–Loveland
 heuristic search in model space (sound but incomplete)
  e.g., min-conflicts-like hill-climbing algorithms
Forward and backward chaining

Horn Form (restricted)

\[ KB = \text{conjunction of Horn clauses} \]

Horn clause =

◊ proposition symbol; or
◊ (conjunction of symbols) \( \Rightarrow \) symbol

E.g., \( C \land (B \Rightarrow A) \land (C \land D \Rightarrow B) \)

Modus Ponens (for Horn Form): complete for Horn KBs

\[
\alpha_1, \ldots, \alpha_n, \quad \alpha_1 \land \cdots \land \alpha_n \Rightarrow \beta
\]

\[
\beta
\]

Can be used with forward chaining or backward chaining. These algorithms are very natural and run in linear time
Forward chaining

Idea: fire any rule whose premises are satisfied in the $KB$, add its conclusion to the $KB$, until query is found

\[
\begin{align*}
P & \Rightarrow Q \\
L \land M & \Rightarrow P \\
B \land L & \Rightarrow M \\
A \land P & \Rightarrow L \\
A \land B & \Rightarrow L \\
A & \\
B &
\end{align*}
\]
function PL-FC-ENTAILS?(KB, q) returns true or false
local variables: count, a table, indexed by clause, initially the number of premises
inferred, a table, indexed by symbol, each entry initially false
agenda, a list of symbols, initially the symbols known to be true

while agenda is not empty do
    p ← POP(agenda)
    unless inferred[p] do
        inferred[p] ← true
        for each Horn clause c in whose premise p appears do
            decrement count[c]
            if count[c] = 0 then do
                if HEAD[c] = q then return true
                PUSH(HEAD[c], agenda)
    return false
Forward chaining example
Forward chaining example
Forward chaining example
Forward chaining example
Forward chaining example
Forward chaining example
Forward chaining example
Forward chaining example
Proof of completeness

FC derives every atomic sentence that is entailed by $KB$

1. FC reaches a fixed point where no new atomic sentences are derived

2. Consider the final state as a model $m$, assigning true/false to symbols

3. Every clause in the original $KB$ is true in $m$
   
   Proof: Suppose a clause $a_1 \land \ldots \land a_k \Rightarrow b$ is false in $m$
   
   Then $a_1 \land \ldots \land a_k$ is true in $m$ and $b$ is false in $m$
   
   Therefore the algorithm has not reached a fixed point!

4. Hence $m$ is a model of $KB$

5. If $KB \models q$, $q$ is true in every model of $KB$, including $m$

   General idea: construct any model of $KB$ by sound inference, check $\alpha$
Backward chaining

Idea: work backwards from the query $q$:
   to prove $q$ by BC,
       check if $q$ is known already, or
       prove by BC all premises of some rule concluding $q$

Avoid loops: check if new subgoal is already on the goal stack

Avoid repeated work: check if new subgoal
   1) has already been proved true, or
   2) has already failed
Backward chaining example

Q → P → M → L → A, B
Backward chaining example
Backward chaining example
Backward chaining example
Backward chaining example
Backward chaining example
Backward chaining example
Backward chaining example
Backward chaining example

Diagram showing the backward chaining process with nodes A, B, L, M, P, and Q.
Backward chaining example
Backward chaining example
Forward vs. backward chaining

FC is **data-driven**, cf. automatic, unconscious processing,
   e.g., object recognition, routine decisions

May do lots of work that is irrelevant to the goal

BC is **goal-driven**, appropriate for problem-solving,
   e.g., Where are my keys? How do I get into a PhD program?

Complexity of BC can be *much less* than linear in size of KB
Resolution

Conjunctive Normal Form (CNF—universal)

*conjunction* of *disjunctions* of *literals* *clauses*

E.g., \((A \lor \neg B) \land (B \lor \neg C \lor \neg D)\)

Resolution inference rule (for CNF): complete for propositional logic

\[
\begin{array}{c}
\ell_1 \lor \cdots \lor \ell_k, \quad m_1 \lor \cdots \lor m_n \\
\ell_1 \lor \cdots \lor \ell_{i-1} \lor \ell_{i+1} \lor \cdots \lor \ell_k \lor m_1 \lor \cdots \lor m_{j-1} \lor m_{j+1} \lor \cdots \lor m_n \\
\end{array}
\]

where \(\ell_i\) and \(m_j\) are complementary literals. E.g.,

\[
\begin{array}{c}
P_{1,3} \lor P_{2,2}, \quad \neg P_{2,2} \\
P_{1,3} \\
\end{array}
\]

Resolution is sound and complete for propositional logic
Conversion to CNF

\[ B_{1,1} \iff (P_{1,2} \lor P_{2,1}) \]

1. Eliminate \( \iff \), replacing \( \alpha \iff \beta \) with \( (\alpha \implies \beta) \land (\beta \implies \alpha) \).

\[ (B_{1,1} \implies (P_{1,2} \lor P_{2,1})) \land ((P_{1,2} \lor P_{2,1}) \implies B_{1,1}) \]

2. Eliminate \( \implies \), replacing \( \alpha \implies \beta \) with \( \neg \alpha \lor \beta \).

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land (\neg (P_{1,2} \lor P_{2,1}) \lor B_{1,1}) \]

3. Move \( \neg \) inwards using de Morgan’s rules and double-negation:

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land ((\neg P_{1,2} \land \neg P_{2,1}) \lor B_{1,1}) \]

4. Apply distributivity law (\( \lor \) over \( \land \)) and flatten:

\[ (\neg B_{1,1} \lor P_{1,2} \lor P_{2,1}) \land (\neg P_{1,2} \lor B_{1,1}) \land (\neg P_{2,1} \lor B_{1,1}) \]
Resolution algorithm

Proof by contradiction, i.e., show $KB \land \neg \alpha$ unsatisfiable

function PL-RESOLUTION($KB, \alpha$) returns true or false

    clauses ← the set of clauses in the CNF representation of $KB \land \neg \alpha$
    new ← {}$
    loop do
        for each $C_i, C_j$ in clauses do
            resolvents ← PL-RESOLVE($C_i, C_j$)
            if resolvents contains the empty clause then return true
            new ← new $\cup$ resolvents
            if new $\subseteq$ clauses then return false
        clauses ← clauses $\cup$ new
Resolution example

\[ KB = (B_{1,1} \iff (P_{1,2} \lor P_{2,1})) \land \neg B_{1,1} \land \alpha = \neg P_{1,2} \]
Summary

Logical agents apply inference to a knowledge base to derive new information and make decisions.

Basic concepts of logic:
- **syntax**: formal structure of sentences
- **semantics**: truth of sentences wrt models
- **entailment**: necessary truth of one sentence given another
- **inference**: deriving sentences from other sentences
- **soundness**: derivations produce only entailed sentences
- **completeness**: derivations can produce all entailed sentences

Wumpus world requires the ability to represent partial and negated information, reason by cases, etc.

Forward, backward chaining are linear-time, complete for Horn clauses.
Resolution is complete for propositional logic.

Propositional logic lacks expressive power.