Distinguishing Abduction and Induction under Intensional Complexity

José Hernández-Orallo
Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació
Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain.
E-mail: `jorallo@dsic.upv.es`.

Ismael García-Varea
Universitat Politècnica de València
Institut Tecnològic d'Informàtica
Camí de Vera 14, E-46071, València, Spain.
E-mail: `ivarea@iti.upv.es`

ECAI'98 Workshop on Abduction and Induction in AI,
Brighton 25th August

Introduction

Semantically, given an evidence C and a background theory T, non-deductive inference tries to obtain A from T and C.

$$\boxed{A \cup T' \models C}$$

Different motivations:

- Purpose: Descriptional / Predictive / Explanatory
- Kind: Enumerative (laws) / Assumptive (facts)
- Justification: Causal / Non-causal

$$\Downarrow$$

Different Paradigms: Enumerative Induction, Explanatory Induction, Best Explanation, Abduction…

Is there an *intrinsical* differentiation of non-deductive inference mechanisms?

Intrinsical Criteria

Syntactical (*How is the hypothesis?*)
- Syntactic Complexity (MDL principle).

Semantical (*What does the hypothesis cover?*):
- 'Informativeness' (Popper) vs. Non-Presumptiveness.
- Generality vs. Specificity.
- Exact-Complete vs. Approximate-Partial.

'Behavioural': (*How does the hypothesis cover the evidence?*)
- Computational (time) Complexity.
- 'Consilience' (Whewell 1847) — 'Common Cause' (Reichenbach 1956) — Coherence (Thagard 1978) *vs.* Separate Covering.
- Intensionality vs. Tolerance of Partial Extensionality.

Abduction

The semantical schema is apparently the same:

$$\boldsymbol{A \cup T \models C}$$

Characterising restrictions (*some of them incompatible*):

- *A* is usually a fact (easy for FOL but not in general).
- *C* should be explained by *A* in the context of *T*. ($A \not\models C$).
- *A* should be likely (the least presumptive, the shortest…).
- *A* should be informative ($A \neq C$, the most presumptive…).
- *A should not* be an uncertain, non-monotonic or probabilistic deduction from *T* (non-nomological abduction).
- *A* must be simple: wrt. inclusion (subset minimality) and syntactically (MDL).

Abduction shares with Induction the most important dilemma between likely *vs.* informative hypotheses.

- We want a hypothesis that should be informative but, at the same time, it should be a *matter of course*.

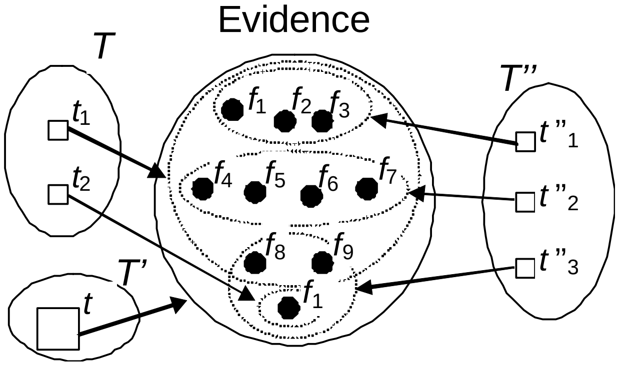
What is a matter of course?

Matter of Course and Exceptions

Intrinsical Exception: *Something we can take apart from a hypothesis so leaving it much simpler wrt. the magnitude of the evidence which would become uncovered.*

EXAMPLE:

- Evidence C= { f_1, f_2, … f_{10} }
- Hypotheses T = { t_1, t_2 }, T’ = {t’’} and T’’ = {t’$_1$, t’$_2$, t’$_3$}
  where T is shorter than T’ and shorter than T’’.



- Is f_{10} a matter of course wrt. T?   wrt. T’?   wrt. T’’?

T’’ shows there are two different (but closely related) notions:

- *T*’ and *T*’’’ are *intensional*. They have no exceptions.
- *T* and *T*’’ are separable. They are not *consilient*.

Intensional Complexity

$\Delta(p) = e$ denotes the no. of exceptions e of a description p.

DEFINITION 2.1. INTENSIONAL COMPLEXITY

The *Intensional Complexity* (IC) of a string x on a bias β:

$$E_\beta(x) = \min \{ l_\beta(p_x) : \Delta(p_x) = 0 \}$$

p_x denotes any program for x in β and $l_\beta(p_x)$ denotes the length of p_x in β.

i.e. the shortest program for x without intrinsic exceptions.

$E(h)$ integrates avoidance of exceptions, consilience and syntactical simplicity because:

- A formal definition of $\Delta(p)$ for any descriptional mechanism requires a general definition of *subprogram*. This must be necessary based on the idea of separation: "*something is separable if the cost of describing the whole is similar to the cost of describing the parts*" which is as well very related to the idea of exception.

The prior $P(h) = 2^{-E(h)}$ could be seen as an adaptation for explanation of the MDL principle ($P(h) = 2^{-K(h)}$).

- *Simplicity is important but secondary.*
- *Nothing is noise or casual, all must be explained. All is intensional. All has a meaning, a cause…*

Consilient Logic Theories

Logic theories or programs composed of Horn rules.
Minimal Herbrand model $M^+(P)$ defined as usual.

DEFINITION 4.1. SEPARABLE PROGRAMS
A program P is n-separable in the partition of *different* programs $\Pi = \{ P_1, P_2, \ldots , P_n \}$ iff

$$M^+(P) = \bigcup_{i=1..n} M^+(P_i) \quad \text{and} \quad \forall_{i=1..n} \; M^+(P_i) \neq \emptyset$$

Additional restrictions (*modes*) of separation:

I. *non-empty*: DEF 4.1
II. *non-subset*: DEF 4.1 $+\forall_{i,j=1..n} (P_i \subseteq P_j \Rightarrow i=j)$.
III. *disjoint*: DEF 4.1 $+ \forall_{i,j=1..n} (Pi \cap Pj = \emptyset)$.
IV. *non-subset <u>model</u>*: DEF 4.1 $+\forall_{i,j=1..n} (M^+(Pi) \subseteq M^+(Pj) \Rightarrow i=j)$.
V. *disjoint <u>model</u>*: DEF 4.1 $+\forall_{i,j=1..n} (M^+(Pi) \cap M^+(Pj) = \emptyset)$.

A theory is consilient iff it is not separable.

- The modes give 5 characterisations of consilient theories.

Example

EXAMPLE:

- P_1= { p(a). q(X) :- r(X). r(a). } is {i-v}separable into Π = {{p(a)} , {q(X) :- r(X). r(a)}}.
- P_2= { q(X) :- r(X). r(b). } is not {i-v}separable.
- P_3= { q(X) :- r(X). p(X) :- r(X). r(a). } is non-subset (model) separable into Π = {{ q(X) :- r(X). r(a)}, {p(X) :- r(X). r(a). }} but it is not disjoint (model) separable.
- P_4= { q(a). p(X) :- q(X). p(a) } is non-subset (model) and disjoint separable into Π = {{ q(a). p(X) :- q(X). }, {p(a)}} but it is not disjoint model separable.
- P_5= { s(X):- p(X), q(b). p(X) :- q(X). t(X):-p(X),q(a) } is non-subset (model) and disjoint separable model into Π = {{ s(X) :- p(X), q(b). p(X) :- q(X) }, { p(X) :- q(X), t(X) :- p(X), q(a) } but it is not disjoint separable.

Problems of non-modularity (I, II, IV):

Problems of fantastic consilient concepts (III, V):

P_1 can be *consiliated* by a fantastic concept *f* into P'_1= { p(a) :- *f*. q(X) :- r(X), *f*. r(a), *f*. *f*. } for iii-iv.

Exception-Free Logic Theories

DEFINITION 4.6. EXCEPTIONS IN A LOGIC PROGRAM

A program P has $e = \text{card}(M^+(P_E))$ c-exceptions generated from P_E, denoted $\Delta_C(P, P_E) = e$, iff there is a partition $P = \{ P_R, P_E \}$ such that:

$$l(\text{P}) - l(\text{P}_R) \geq (1 / c) \cdot [l(M^+(P)) - l(M^+(P_R))]$$

where l denotes any syntactical measure of length.

If P_E is not specified, $\Delta_C(P) = \max \{ e \mid \Delta_C(P, P_E) = e \}$

Fixing l and an exception-degree c (usually $c = 1$), a theory P is said to be exception-free iff $\Delta_C(P) = 0$

Pragmatics:

- The modes give 5 characterisations of intensional (exception-free) theories.
- Mode ii and c=1 allow modular programs and avoid fantastic concepts.
- For instance, P = { p(X). q(X) } for evidence { p(a), p(b), p(e), q(a), q(d), q(e), q(f) } is separable but it has no exceptions.

Exceptions and Abduction

$$\boldsymbol{A \cup T \models C}$$

A must be a matter of course. It cannot be an exception wrt. to T. $\Rightarrow$ Apply DEF. 4.6 and choose $P_R = T$.

EXAMPLE:

Program T = { p.
lawn-wet :- rain.
lawn-wet :- sprinkler-on. }

Observation C = { lawn-wet },
and the following short explanations:
$A_1 = C$, A_2 = {rain}, A_3 = {sprinkler-on}, A_4 = {lawn-net :- p}

- A_1 is an exception because $l(A_1 \cup T) - l(T) = l(A_1) \geq l(M^+(T) + C)) - l(M^+(T)) = l(C)$.
- A_2 (and A_3) are not because we have $l(A_2 \cup T) - l(T) = l(A_2) < l(\mathrm{M}^+(T) + C + A_2) - l(M^+(T)) = l(C + A_2)$.
- A_4 is also an exception because $l(A_4) \geq l(M^+ (T) + C)) - l(M^+(T)) = l(C)$, so it is not a valid explanation.

Incremental Setting

Knowledge Acquisition and Revision

A theory T is constructed as the data suggest.

Each time a new observation C is perceived, there are three possible situations:

- **Prediction Hit**. The observations are covered without more assumptions, i.e., $T \models C$. The theory is reinforced.
- **Novelty**. The observation is uncovered but consistent with the theory T , i.e., $T \not\models C$ and $T \cup C \not\models \square$. Here, the possible actions are:
  1. *Extension*: T can be extended with a good explanation,
  2. *Revision*: T can be modified if a coherent explanation cannot be found,
  3. *Patch*: left it as an intrinsical exception, or
  4. *Rejection*: ignored.
- **Anomaly**. The observation is inconsistent with the theory T, i.e., $T \not\models C$ and $T \cup C \models \square$. In this case, we have three possibilities: *revision*, *patch* or *rejection*.

Reinforcement

Further detail on the relation hypothesis ⩽ evidence:

DEFINITION 7.1. PURE REINFORCEMENT

The pure reinforcement $\rho\rho(r)$ of a rule r from a theory T wrt. to some given observation $C = \{ c_1, c_2, \ldots, c_n \}$ is computed as the number of proofs of c_i where r is used. If there are more than one proof for a given c_i, *all* of them are reckoned. In the same proof, a rule is computed once.

DEFINITION 7.2. NORMALISED REINFORCEMENT

$$\rho(r) = 1 - 2^{-\rho\rho(r)}.$$

Properties:

- The most reinforced theory is not the shortest one.
- Redundancy does not imply a loss of reinforcement ratio.
- Measure is wrt. the *theory* → fantastic concepts.

DEFINITION 7.3. REINFORCEMENT WRT. THE DATA

The *course* $\chi(f)$ of a given fact f wrt. to a theory is computed as the product of all the reinforcements $\rho(r)$ of all the rules r used in the proof of f. If a rule is used more than once, it is computed once. If f has more than one proof, we select the greatest course.

Characteristics of Reinfocement

- Redundancy is possible, although MDL usually ensures a good mean course ratio.
- However, theorem 7.1 shows that the use of *fantastic* concepts cannot increase *artificially* the courses.

☝Advantages:

- Reinforcement is easy to compute and allows a flexible evaluation of a theory and the data it covers.
- It provides a measure of the predictive accuracy or assumption feasibility.
- It works for evidences with noise.

☟Drawbacks:

- Theories cannot be evaluated for infinite evidences.

Selection Criteria:

- The Most Reinforced One: The greatest *mean* ($m\chi$)of the courses of all the data presented so far.
- More Compensated: a *geometric mean* instead.
- Intensional: all facts should have a course value greater than the mean divided by a constant (no exceptions).
- Consilience can be better studied: a theory is well-separable if $m\chi$ is not decreased after separation.

Examples:

1) E = { p(a), p(b), p(e), q(a), q(d), q(e), q(f) }

P = { p(X) : ρ = 0.875

q(X) : ρ = 0.9375 } $m\chi(E,P) = 0.90625$

P_1 = { p(X) : ρ = 0.875 }

P_2 = { q(X) : ρ = 0.9375 } $m\chi(E, P_1 \oplus P_2) =$ 0.90625

2) E = { q(a),p(a),¬r(a), q(b),p(b),r(b), q(c),¬p(c), ¬q(d),¬q(e) }

P_a = { p(a) : ρ = 0.75

r(b) : ρ = 0.875

q(X) :- p(X) : ρ = 0.75

p(X) :- r(X) : ρ = 0.875

q(c) : ρ = 0.5 } $m\chi(E,P_a) = 0.6393$ (low)

$P_{MDL} = E^+$ $m\chi(E,P_{MDL}) = 0.5$ (very low)

Abduction is possible with P_a:

Evidence: q(f)

Possible Assumptions:

q(f)? $m\chi(E,P_a \cup \{ q(f) \}) = 0.619$

p(f)? $m\chi(E,P_a \cup \{ p(f) \}) = 0.627$

r(f)? $m\chi(E,P_a \cup \{\, \mathrm{r(f)}\, \}\,) = 0.657$

Long Example (1 from 3)

Incremental learning session:

♦ Background theory

B = { s(a,b), s(b,c), s(c,d) }

we observe the evidence

E = { e_1^+: r(a,b,c),
e_2^+: r(b,c,d),
e_3^+: r(a,c,d),
e_1^-: ¬r(b,a,c),
e_2^-: ¬r(c,a,c) }

Hypotheses:

P_1={r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.875$

P_2={r(X,c,Z) : ρ = 0.75
r(a,Y,Z) : ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.75$

P_3={r(X,Y,Z) :- s(X,Y) : ρ = 0.75
r(X,Y,Z) :- s(Y,Z) : ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= 0.875$

P_4={r(X,Y,Z):-t(X,Y),t(Y,Z): ρ=0.875
t(X,Y):-s(X,Y): ρ = 0.875
t(X,Y):-s(X,Z),t(Z,Y): ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= 0.7656, \chi(e_3^+)= 0.3828$

P_5={r(X,Y,Z) :- t(X,Y) : ρ = 0.875
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y):-s(X,Z),t(Z,Y): ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= 0.7656, \chi(e_3^+)= 0.3828$

At this moment, P_1 and P_3 are the best options by far.
P_4 and P_5 seem fantastic theories according to the evidence

Long Example (2 from 3)

♦ e_4^+ = r(a,b,d) is observed.

P_1 does not cover e_4^+ and it is patched to:

P_{1a}'={r(X,Y,Z):-s(Y,Z) : ρ = 0.875
r(a,b,d) : ρ = 0.5}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+) = 0.875$, $\chi(e_4^+) = 0.5$
Mean = 0.78, GeoMean = 0.76

P_{1b}'={r(X,Y,Z):-s(Y,Z) : ρ = 0.875
r(X,Y,d) : ρ = 0.875 }
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+) = 0.875$

P_2' is reinforced
P_2'={r(X,c,Z) : ρ = 0.75.
r(a,Y,Z) : ρ = 0.875}
$\chi(e_1^+) = 0.875$, $\chi(e_2^+) = 0.75$, $\chi(e_3^+)= \chi(e_4^+) = 0.875$

P_3' is reinforced
P_3'={r(X,Y,Z): s(X,Y) : ρ = 0.875.
r(X,Y,Z):-s(Y,Z): ρ = 0.875}
$\chi(e_1^+)= \chi(e_2^+)= \chi(e_3^+)= \chi(e_4^+)= 0.875$

P_4' is reinforced.
P_4'={r(X,Y,Z):-t(X,Y),t(Y,Z):ρ=0.9375
t(X,Y) :- s(X,Y) : ρ = 0.9375
t(X,Y) :- s(X,Z), t(Z,Y):ρ = 0.75}
$\chi(e_1^+)= \chi(e_2^+)= 0.8789$,
$\chi(e_3^+)= \chi(e_4^+) = 0.6592$
Mean= 0.77, GeoMean = 0.76

P_5' is slightly reinforced
P_5'={r(X,Y,Z):-t(X,Y):ρ = 0.9375.
t(X,Y) :- s(X,Y) : ρ = 0.875
t(X,Y):-s(X,Z),t(Z,Y): ρ= 0.5}

At this moment, P_{1b}' and P_3 are the best options. Now P_4 seems less fantastic.

Long Example (3 from 3)

♦ We add $e_3^- = \neg r(a,d,d)$

P_{1a}' remains the same.

P_{1b}' and P_{2a}' are inconsistent. The following two theories could also be 'patches' for them:

$P_{2a}' = \{r(X,c,Z) : \rho = 0.75.$
$r(X,b,Z) : \rho = 0.75\}$
$\chi(e_1^+) = \chi(e_2^+) = \chi(e_3^+) = \chi(e_4^+) = 0.75$

$P_{2b}' = \{r(X,Y,Z) :- e(Y) : \rho = 0.9375.$
$e(b) : \rho = 0.75$
$e(c) : \rho = 0.75\}$
$\chi(e_1^+) = \chi(e_2^+) = \chi(e_3^+) = \chi(e_4^+) = 0.7031$

P_3' and P_4' remain the same and P_5' seem to be inconsistent.

♦ We add $e_5^+ = r(a,d,e)$

P_{1a}', P_{2a}', P_{2b}' can only be patched with e_5^+ as an exception and not abduction is possible.

P_3' has abduction as a better option.

$P_3'' = \{ s(d,e) : \rho = 0.5$
$r(X,Y,Z) :- s(X,Y) : \rho = 0.875$
$r(X,Y,Z) :- s(Y,Z) : \rho = 0.9375\}$
$\chi(e_1^+) = \chi(e_2^+) = \chi(e_3^+) = 0.9375,$
$\chi(e_4^+) = 0.875, \chi(e_5^+) = 0.46875$
Mean= 0.831, GeoMean = 0.805

P_4' makes the same abduction

$P_4'' = \{ s(d,e) : \rho = 0.5$

$r(X,Y,Z) :- t(X,Y), t(Y,Z) : \rho = 0.969$
$t(X,Y) :- s(X,Y) : \rho = 0.96875$
$t(X,Y) :- s(X,Z), t(Z,Y) : \rho = 0.875\}$
$\chi(e_1^+) = \chi(e_2^+) = 0.9385,$
$\chi(e_3^+) = \chi(e_4^+) = 0.8212, \chi(e_5^+) = 0.4106$
Mean= 0.786, GeoMean = 0.754

- The example illustrates that as soon as a theory gains some solidity, abduction can be applied.

Proposed Taxonomy

- *Descriptional (or Enumerative) Induction*: uses background knowledge as a help but it has no expectancy of the source to conciliate (and no restriction either), so a hypothesis is constructed as the data suggest (according to a prior). There may be noise: exceptions are tolerated.
- *Explanatory Induction*: looks for *more informative* theories instead of the most probable. Exceptions are not allowed, because the hypothesis must explain all the data.
- *Abduction*: assumptions (hypothesis that are usually facts) should be a "matter of course" wrt. the background knowledge, i.e. not only consistency but also consilience is required.

The difference between enumerative and explanatory induction is the *intensionality* of the hypothesis (avoidance of exceptions).

The *subtle* distinction between Explanatory Induction and Abduction resides in that, for the latter, $A \cup T$ must be consilient, and it is only possible when T has more relative importance and validation wrt. to A.

Conclusions

- Syntactic and Semantic considerations are not sufficient to distinguish between induction and abduction.
- The relation between the hypothesis and the evidence (i.e. *how the hypothesis covers the data*) allow further insight in the evaluation and character of the hypotheses.

⌦ Intensionality and Presence of Noise: there are *acceptable* explanations in the presence of noise.

  * We can use the intrinsic degree or percentage of exceptions $\Delta_c(p) / n$ being n the number of examples. If we know the noise ratio ε, the hypotheses should observe $\Delta_c(h) / n = \varepsilon$.

Current and Future work

- Evaluating in practice these *intensional* principles in inductive systems [Hernandez-Orallo & Ramirez-Quintana 1998].
- Integrate reinforcement propagation for deductive inference and negative evidence. Relate with non-monotonic reasoning frameworks.
- Extend the incremental knowledge construction setting to interactive frameworks (query learning or actions and reward) and the common view of *reinforcement learning*.