Problem Statement

One big open problem in belief revision, dynamic epistemic logic, and machine learning is to develop the "model theory of learning," i.e.

Question 1. What axioms capture important properties of learning (e.g. induction, no forgetting, no hallucinations, robustness)? What do models of these properties look like? Are there "correspondences" between axioms and properties of models, as there often are in modal logic?

A lot of work has already been done on this — see the two very relevant papers [1, 2]. In particular, [1] shows that, given a stream of true data, announcement and lexicographic upgrade converge to the full truth (whereas conservative upgrade does not). And if the stream has finitely many false data, only lexicographic upgrade converges to the full truth (of the three).

I was reading the second volume of [3] recently. In it, George Pólya makes some empirical observations about the role of guessing and induction in mathematics, and then tries to formalize these as logical rules. I got interested in this because I noticed that his rules of induction capture axioms for *plausibility upgrade / belief revision* that I believe have been overlooked by work on DEL and belief revision.

For the rest of this note, I want to formalize Pólya's inductive rules as DEL axioms and explore the models that satisfy them. This is a bit of a detour from the question above, but I've been having fun thinking about it (and I hope you do too)!

Pólya's Stepwise Induction

Pólya charts about 16 different rules in his book, but for now I'm interested in the two that relate to induction. First, let's look at his Stepwise Induction rule

$$\frac{P \to Q \qquad Q \text{ true}}{P \text{ more plausible}}$$

Pólya gives this rule the following reading: If we discover that some consequence Q of P is true, then P becomes more plausible. He refers to this as just "induction"; I prefer to call it *stepwise* induction because a modern understanding of induction is *convergence to the truth*, and this rule says what ought to happen in a single discovery step rather than at some point of convergence.

Pólya wrote this 30 years before logics for plausibility upgrade were developed. But with the hindsight we have today, it's clear that his rule expresses a DEL constraint for plausibility upgrade.

Suppose we have a plausibility model $\mathcal{M} = \langle W, R, \leq, V \rangle$. Let $[P]$ be the action "make P more plausible," which updates $\mathcal{M}$ to $\mathcal{M}^{\star}_{P}$. Let $[!P]$ be the action "discover that P is true, which updates $\mathcal{M}$ to $\mathcal{M}^{!}_{P}$ (we could do this by conditionalization / public announcement).

To make things concrete, let's consider the full language

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Diamond\varphi \mid \mathbf{K}\varphi \mid \mathbf{B}^{\varphi}\psi \mid [!P]\varphi \mid [P]\varphi$$

with the semantics

$$\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash \Diamond\varphi & \text{iff} & \text{there exists } u \text{ with } \mathcal{M}, u \Vdash \varphi \\
\mathcal{M}, w \Vdash \mathbf{K}\varphi & \text{iff} & \text{for all } u \text{ with } wRu, \mathcal{M}, u \Vdash \varphi \\
\mathcal{M}, w \Vdash \mathbf{B}^{\varphi}\psi & \text{iff} & \mathsf{best}_{\leq}([\![\varphi]\!]) \subseteq [\![\psi]\!] \\
\mathcal{M}, w \Vdash [!P]\varphi & \text{iff} & \mathcal{M}^{!}_{P}, w \models \varphi \\
\mathcal{M}, w \Vdash [P]\varphi & \text{iff} & \mathcal{M}^{\star}_{P}, w \models \varphi
\end{array}$$

where $[\![\varphi]\!] = \{u \mid \mathcal{M}, u \Vdash \varphi\}$, and $\mathsf{best}_{\leq}(S) = \{u \in S \mid u \text{ is } \leq\text{-minimal over } S\}$.

We can't express the rule in DEL as-stated, since $[P]$ and $[!P]$ can't occur by themselves. But we can rephrase the rule in terms of the *effects* that the updates have: The rule says that any effect of making P more plausible is also an effect of discovering that Q is true. More formally, if $\mathcal{M}, w \Vdash P \to Q$ then

$$\mathcal{M}^{\star}_{P}, w \Vdash \varphi \text{ implies } \mathcal{M}^{!}_{Q}, w \Vdash \varphi$$

And this holds iff $\mathcal{M}, w$ satisfies

$$(P \to Q) \to ([P]\varphi \to [!Q]\varphi) \quad \text{(STEP-IND)}$$

The second rule I'd like to explore is Shaded Stepwise Induction ("shaded" is Pólya's terminology)

$$\frac{P \to Q \qquad Q \text{ more plausible}}{P \text{ more plausible}}$$

We can do a similar analysis on this, and express it in DEL as

$$(P \to Q) \to ([P]\varphi \to [Q]\varphi) \quad \text{(SHADED-STEP-IND)}$$

I want to ask the same questions from before, but now zoomed-in at these two axioms. Specifically,

Question 2. What do satisfying models of (STEP-IND) and (SHADED-STEP-IND) look like? I'll be happy to find just one for each, but I'd like to completely axiomatize them if possible.

Question 3. What is the relationship between (STEP-IND) and (SHADED-STEP-IND)? Is one 'stronger' than the other, and in what sense?

Question 4. What is the relationship between *stepwise* induction and induction understood as convergence to the truth? Does stepwise induction imply regular induction (as the name suggests)? Or is this notion of stepwise induction irrelevant to regular induction?

Progress So Far

A Model for Shaded Stepwise Induction

Definition 1. Given $\mathcal{M} = \langle W, R, \leq, V\rangle$, let the update $\mathcal{M}^\star_P = \langle W, R, \leq^\star_P, V\rangle$, where $\leq^\star_P$ is reordered according to the rule:

> Put all states where *some* (non-$\top$) consequence of P holds in front of all the other states; the order within the two groups is kept the same.

Note that this is a variation on Lexicographic upgrade.

Lemma 1. For all formulas P, Q and all sets S, $\mathsf{best}_{\leq^\star_Q}(S) \subseteq \mathsf{best}_{\leq^\star_P}(S)$.

Proof. We have two cases:

- No state $w \in S$ satisfies a consequence of Q. Note that this also implies that no state satisfies a consequence of P, since TODO

□

Theorem 2. (SHADED-STEP-IND) is sound for this update.

Proof. By induction on φ. The interesting case is conditional belief $\mathbf{B}^\varphi\psi$, since it is the only case affected by the upgrade. Suppose $\mathcal{M}, w \Vdash P \to Q$, and $\mathcal{M}, w \Vdash [P]\mathbf{B}^\varphi\psi$. So $\mathsf{best}_{\leq^\star_P}(\llbracket\varphi\rrbracket) \subseteq \llbracket\psi\rrbracket$. By Lemma 1 we have

$$\mathsf{best}_{\leq^\star_Q}(\llbracket\varphi\rrbracket) \subseteq \mathsf{best}_{\leq^\star_P}(\llbracket\varphi\rrbracket) \subseteq \llbracket\psi\rrbracket$$

And so $\mathcal{M}, w \Vdash [Q]\mathbf{B}^\varphi\psi$. □

So we have an update that models (SHADED-STEP-IND). My goal will now be to completely axiomatize this update.

Proposition 3. We can completely characterize the effect this update has on best. For all $S \subseteq W$,

$$\mathsf{best}_{\leq^\star_P}(S) = \begin{cases} \mathsf{best}_{\leq}(S \cap \{w \mid \exists X \not\equiv \top \text{ such that } \mathcal{M}, w \Vdash P \to X \text{ and } \mathcal{M}, w \Vdash X\}) \\ \qquad \text{if } S \cap \{w \mid \exists X \not\equiv \top \text{ such that } \mathcal{M}, w \Vdash P \to X \text{ and } \mathcal{M}, w \Vdash X\} \neq \emptyset \\ \mathsf{best}_{\leq}(S) \quad \text{otherwise} \end{cases}$$

Proof. Notice that this expression is exactly the characterization of lexicographic upgrade

$$\mathsf{best}_{\leq^\Uparrow_P}(S) = \begin{cases} \mathsf{best}_{\leq}(S \cap \llbracket P\rrbracket) & \text{if } S \cap \llbracket P\rrbracket \neq \emptyset \\ \mathsf{best}_{\leq}(S) & \text{otherwise} \end{cases}$$

except instead of upgrading the set

$$\llbracket P\rrbracket = \{w \mid \mathcal{M}, w \Vdash P\}$$

we upgrade

$$\{w \mid \exists X \not\equiv \top \text{ such that } \mathcal{M}, w \Vdash P \to X \text{ and } \mathcal{M}, w \Vdash X\}$$

From here, the proof is identical to the proof for lexicographic upgrade. □

Discussion. Consider the expression

$$\exists X \not\equiv \top \text{ such that } \mathcal{M}, w \Vdash P \to X \text{ and } \mathcal{M}, w \Vdash X$$

If we could express this in DEL, we would have a complete axiomatization for the upgrade (just modify the reduction axioms for lexicographic upgrade). This has an undeniably modal flavor to it; think of $P \to X$ as an accessibility relation over formulas. It doesn't seem to be expressible in our language so far, so let's go ahead and define it.

Definition 2. Let ◎ be a new operator in our language, with semantics given by

$$\mathcal{M}, w \Vdash ◎P \text{ iff } \exists \text{ formula } X \not\equiv \top \text{ such that } \mathcal{M}, w \Vdash P \to X \text{ and } \mathcal{M}, w \Vdash X$$

We read $◎P$ as "some consequence of P holds." ◎ is an existential modality; we define the dual $[◎]P \leftrightarrow \neg◎\neg P$ ("all consequences of P hold"). (The ◎ symbol evokes an image of the winding paths we may take in finding consequences of P.)

Proposition 4. The following are sound:

1. If $\vdash P$ then $\vdash [◎]P$ (NECESS)
2. $(P \to Q) \to (◎Q \to ◎P)$ (ANTITONE)
3. $P \to ◎P$ (REFL)

Proof. (1) is easy: If P holds in all $\mathcal{M}, w$, then so do all of its (nontrivial) consequences. (3) is also easy, since P is a consequence of itself.

For (2), suppose $\mathcal{M}, w \Vdash P \to Q$ and $\mathcal{M}, w \Vdash ◎Q$. Then $\mathcal{M}, w \Vdash X$ for some $X \not\equiv \top$ such that $\mathcal{M}, w \Vdash Q \to X$. But then by transitivity of consequence (X is a consequence of Q, which is a consequence of P), $\mathcal{M}, w \Vdash P \to X$, which gives us $\mathcal{M}, w \Vdash ◎P$. □

Corollary 5. Here are some interesting consequences of (ANTITONE) and (REFL). Notice that they are flipped from their usual □ presentation.

1. $◎P \wedge ◎Q \to ◎(P \wedge Q)$ (M)
2. $◎(P \vee Q) \to ◎P \vee ◎Q$ (OR)
3. $◎◎P \leftrightarrow ◎P$ (TRANS)

Proof. For (1), since $\vdash P \wedge Q \to P$, by (ANTITONE) we have $\vdash ◎P \to ◎(P \wedge Q)$. But then we have $\vdash ◎P \wedge ◎Q \to ◎(P \wedge Q)$. (2) is similar: Since $\vdash P \to P \vee Q$, by (ANTITONE) we have $\vdash ◎(P \vee Q) \to ◎P$. Consequently, $\vdash ◎(P \vee Q) \to ◎P \vee ◎Q$.

Now consider (3). The backwards direction is just (REFL). As for the other direction, (REFL) gives us $\vdash P \to ◎P$. Applying (ANTITONE) we get $\vdash ◎◎P \to ◎P$. □

Importantly, ◎ is *not* a normal modality — the converse of (M), $◎(P \wedge Q) \to (◎P \wedge ◎Q)$, is not sound. Surprisingly, ◎ is completely axiomatized by (NECESS), (ANTITONE), and (REFL).

Lemma 6. (Existence Lemma). Suppose Δ is a maximally consistent set with $◎\varphi \in \Delta$. Then there is a formula $X \not\equiv \top$ of complexity less than or equal to φ with $\varphi \to X \in \Delta$ and $X \in \Delta$.

Proof. Suppose for contradiction that there is no such X, i.e. for all X of complexity less than or equal to φ, if $\varphi \to X \in \Delta$ then $X \notin \Delta$. Let's now construct the formula

$$\psi = \bigvee \{X \not\equiv \top \mid X \text{ has complexity} \leq \varphi \text{ and } \varphi \to @X \in \Delta\}$$

Intuitively, this formula says "some consequence of some consequence of φ holds." Is it well-defined? Well, by (REFL) $\varphi \to @\varphi \in \Delta$, so the set is nonempty. Moreover, since the complexity of these formulas is bounded there are only finitely many of them. So this formula is well-defined.

First, since $@\varphi \in \Delta$ and Δ is maximally consistent, $\psi \to @\varphi \in \Delta$ (by $\to$-introduction: if we suppose $\psi \in \Delta$ then trivially $@\varphi \in \Delta$).

Second, I claim that $\neg @\psi \in \Delta$. Why? Let $\psi = X_1 \vee \ldots \vee X_k$. For each X_i occurring in ψ, $(\varphi \to @X_i) \in \Delta$. By our very first assumption, every $@X_i \notin \Delta$. So $@X_1 \vee \ldots \vee @X_k \notin \Delta$. But then by (OR),

$$@\psi = @(X_1 \vee \ldots \vee X_k) \notin \Delta$$

And since Δ is maximally consistent, we have $\neg @\psi \in \Delta$.

Let's put everything together. Since $\psi \to @\varphi \in \Delta$, by (ANTITONE) we have $@@\varphi \to @\psi \in \Delta$. (REFL) gives us $@\varphi \to @@\varphi$, and then by modus ponens we have $@\varphi \to @\psi \in \Delta$. But since $\neg @\psi \in \Delta$, by contraposition $\neg @\varphi \in \Delta$. But $@\varphi \in \Delta$, so this contradicts the fact that Δ is consistent. □

Proposition 7. Suppose we build the canonical model $\mathcal{M}^c$ using the usual maximally-consistent set construction for the base logic of $\Diamond\varphi$, $\mathbf{K}\varphi$, $\mathbf{B}^\varphi\psi$. The @-case of the truth lemma holds, i.e. for all maximally consistent Δ,

$$\mathcal{M}^c, \Delta \Vdash @\varphi \quad \text{iff} \quad @\varphi \in \Delta$$

Proof. ($\to$) Suppose $\mathcal{M}^c, \Delta \Vdash @\varphi$. So $\mathcal{M}^c, \Delta \Vdash X$ for some $X \not\equiv \top$ such that $\mathcal{M}^c, \Delta \Vdash \varphi \to X$. Now apply the inductive hypothesis to both $\varphi \to X$ and X to get $\varphi \to X \in \Delta$ and $X \in \Delta$. From $X \in \Delta$ and (REFL) we have $@X \in \Delta$. And from $\varphi \to X \in \Delta$, $@X \in \Delta$, and (ANTITONE), we have $@\varphi \in \Delta$.

($\leftarrow$) Now suppose $@\varphi \in \Delta$. By Lemma 6 there is a formula $X \not\equiv \top$ of complexity less than or equal to φ with $\varphi \to X \in \Delta$ and $X \in \Delta$. Since its complexity is less than or equal to φ, we can apply the inductive hypothesis to $\varphi \to X$ and X. So $\mathcal{M}^c, \Delta \Vdash \varphi \to X$ and $\mathcal{M}^c, \Delta \Vdash X$. By the semantics of @, $\mathcal{M}^c, \Delta \Vdash @\varphi$. □

Theorem 8. Assuming model building for the logic $\{\Diamond\varphi, \mathbf{K}\varphi, \mathbf{B}^\varphi\psi\}$, given any consistent Γ over $\{\Diamond\varphi, \mathbf{K}\varphi, \mathbf{B}^\varphi\psi, @\varphi\}$ we can build a model and point $\mathcal{M}, w \models \Gamma$.

Proof. This proof is totally standard:

First, extend Γ to maximally consistent $\Gamma \supseteq \Gamma$ in the usual way, and let $\mathcal{M}^c$ be the canonical model for the logic $\{\Diamond\varphi, \mathbf{K}\varphi, \mathbf{B}^\varphi\psi\}$ (no need to modify it for @). By the Truth Lemma, for all $\varphi \in \Delta$, $\mathcal{M}^c, \Delta \Vdash \varphi$. So in particular, for all $\varphi \in \Gamma$, $\mathcal{M}^c, \Delta \Vdash \varphi$. □

Corollary 9. Assuming completeness for the logic $\{\Diamond\varphi, \mathbf{K}\varphi, \mathbf{B}^\varphi\psi\}$, the logic $\{\Diamond\varphi, \mathbf{K}\varphi, \mathbf{B}^\varphi\psi, @\varphi\}$ is also complete.

Proof. This proof is also standard:

Suppose contrapositively that $\Gamma \nvdash \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is consistent. We can then build a model $\mathcal{M}$ and point w such that $\mathcal{M}, w \Vdash \Gamma \cup \{\neg\varphi\}$. But then $\mathcal{M}, w \Vdash \Gamma$ yet $\mathcal{M}, w \nVdash \varphi$, which is what we wanted to show. □

Theorem 10. The upgrade defined above is completely axiomatized by the reduction axioms (over DEL + ◎):

$$\begin{array}{lcl}
[P]p & \leftrightarrow & p \\
[P]\neg\varphi & \leftrightarrow & \neg[P]\varphi \\
[P](\varphi \wedge \psi) & \leftrightarrow & [P]\varphi \wedge [P]\psi \\
[P]\Diamond\varphi & \leftrightarrow & \Diamond[P]\varphi \\
[P]◎\varphi & \leftrightarrow & ◎[P]\varphi \\
[P]\mathbf{K}\varphi & \leftrightarrow & \mathbf{K}[P]\varphi \\
[P]\mathbf{B}^{\varphi}\psi & \leftrightarrow & (\Diamond(◎P \wedge [P]\varphi) \wedge \mathbf{B}^{◎P \wedge [P]\varphi}[P]\psi) \\
 & & \vee(\neg\Diamond(◎P \wedge [P]\varphi) \wedge \mathbf{B}^{[P]\varphi}[P]\psi)
\end{array}$$

Proof. TODO □

Things To Try

1. go through mathreviews and google scholar to make sure this hasn't been done yet
2. Prove that ◎ is not expressible in the base language (it is a genuine language extension).
3. Find a model satisfying (STEP-IND) (do a similar analysis of this upgrade).
4. Think about Question 3: What is the relationship between (STEP-IND) and (SHADED-STEP-IND)?
5. Think about Question 4: Do either of these axioms gurantee convergence to the truth, under reasonable assumptions?

References

[1] Alexandru Baltag, Nina Gierasimczuk, and Sonja Smets. "Truth-tracking by belief revision". In: *Studia Logica* 107 (2019), pp. 917–947.

[2] Alexandru Baltag et al. "A dynamic logic for learning theory". In: *Journal of Logical and Algebraic Methods in Programming* 109 (2019), p. 100485.

[3] George Polya. *Mathematics and plausible reasoning: Induction and analogy in mathematics*. Vol. 2. Princeton University Press, 1954.