A General Theory of (Identification in the) Limit and Convergence (to the Truth)

Hanti Lin

UC Davis
ika@ucdavis.edu
February 5, 2017

Abstract

I propose a new definition of identification in the limit (also called convergence to the truth), as a new success criterion that is meant to complement, rather than replacing, the classic definition due to Gold (1967). The new definition is designed to explain how it is possible to have successful learning in a kind of scenario that Gold's classic account ignores—the kind of scenario in which the entire infinite data stream to be presented incrementally to the learner is *not* presupposed to completely determine the correct learning target. From a purely mathematical point of view, the new definition employs a convergence concept that generalizes net convergence and sits in between pointwise convergence and uniform convergence. Two results are proved to suggest that the new definition provides a success criterion that is by no means weak: (i) Between the new identification in the limit and Gold's classic one, neither implies the other. (ii) If a learning method identifies the correct target in the limit in the new sense, any U-shaped learning involved therein has to be redundant and can be removed while maintaining the new kind of identification in the limit. I conclude that we should have (at least) two success criteria that correspond to two senses of identification in the limit: the classic one and the one proposed here. They are complementary: meeting any one of the two is good; meeting both at the same time, if possible, is even better.

Keywords: Identification in the Limit, Convergence to the Truth, Language Learning, Enumerative Induction, Uniform Convergence, Net Convergence

1 Introduction

The goal of this paper is to find a new definition of identification in the limit—a new success criterion that is meant to complement, rather than replacing, the classic definition due to Gold (1967). Let me begin with some motivation. Theoretical computer science has a tradition:

(1) An important part of theoretical computer science is basically the science of *problem solving*; in this science, we define a number of *success criteria* for problem solving.

(2) If a difficult problem cannot be solved by meeting a high success criterion, we try to find out whether it can be solved by meeting a success criterion that is lower—or at least *not as high* if success criteria are partially ordered.

(3) If there is an interesting, tough problem that cannot be solved by meeting some success criterion we have defined, we try to find out whether it is possible to define a new criterion that is low enough to be met for solving the tough problem but, simultaneously, still high enough to deserve to be called a *success* criterion

The kind of situation described by (3) has happened a number of times in the history of theoretical computer science. Classic examples in learning theory include the following: When encountering an interesting problem that cannot be solved by a *decision* procedure, i.e. by an *exact* learning method, Valiant (1984) defined a new success criterion called *probable approximate correctness*. In a similar situation, Gold (1967) defined a new success criterion called *identification in the limit*. I suspect that we are now in a similar situation, too. Let me illustrate with an example.

For a simplest example I have in mind, consider the following two languages to be learned. The simpler language contains only one string, 0; and the slightly more complex language contains two strings, 0 and 1. The learner is wondering which language of the two is correct. Whichever is correct, she is to be presented with a sequence of strings that is "sound" and "complete" with respect the correct language—"sound" in the sense of containing *only* the strings in the correct language, and "complete" in the sense of containing *all* of them. The presentation of such a sequence of strings will be incremental, one string at a time. We know that there is no decision procedure for identifying the correct target. But, given that *identification with decision* is unachievable, Gold tells us not to worry and proposes a lower—but still very desirable—success criterion: *identification in the limit*, which requires that, whichever language in the set under consideration is correct, and whichever sound and complete sequence of strings in the correct language

is to be presented incrementally to the learner, the learner's conjecture will eventually converge to the correct language (if, of course, the learner lives long enough).

Call the above problem the **easy language learning problem**, because it is easy to find a learning method that solves this problem by meeting Gold's success criterion. Here is an example: "Conjecture the simpler language when you haven't observed 1; otherwise, conjecture the more complex language." So far so good. But what if the learning problem is made harder?

To be more specific, what if the sequence to be presented to the learner is only required to be sound and allowed to be complete or incomplete—what if the teacher is allowed to be somewhat malicious (although not required to be so)? Call this the **hard language learning problem** because, in this case, no learning method succeeds by Gold's criterion of identification in the limit. Then what to do? I tend to think that this hard problem is very interesting. (How interesting is it? More on this below.) If so, by tradition (1)-(3), we should try to look for an alternative to Gold's success criterion. Furthermore, the task of *at least trying* to do so is not daunting at all. Indeed, Gold only employs the most familiar concept of limit and convergence: convergence of a sequence. But in analysis and topology mathematicians have already worked out more elaborated concepts of convergence (and put them to good work):

- convergence of a *sequence* can be generalized to convergence of a so-called *net*;
- convergence of a sequence of functions comes in two flavors: *pointwise* convergence and *uniform* convergence.

As we will see in this paper, Gold's success criterion basically requires that the learner's convergence to the correct target be "pointwise" and "sequence-like". And, if I am right, it has a very natural variant, which requires that the learner's convergence be "semi-uniform" and "generalized-net-like". Once this idea is developed in rigorous mathematical terms (definition 18), I will be able to show that this new criterion of identification in the limit is achievable for the hard language learning problem (proposition 19). Two further results are proved to suggest that this new criterion is still high and demanding enough to deserve to be called a success criterion: First, the new criterion is neither stronger nor weaker than Gold's classic criterion (theorem 21); second, if the learner can ever solve the problem in the limit in the new sense, then in principle she can does it in pretty much the same way without U-shaped learning (theorem 24). The conclusion to be drawn is that the new success criterion and Gold's classic one are complementary: meeting any one of

them is good; meeting both at the same time is even better. In the unfortunate event that only one can be met, let us meet at least that one.

But why is the hard language learning problem is interesting? Let me provide two reasons. The first reason is short: problems with possibly adversarial and malicious elements have been the focus in the online model of machine learning. Any reason provided therein can, and should, be carried over to the the language-learning model.

The second reason is longer and more important. The defining feature of the hard learning problem does not have to be interpreted as something adversarial or malicious. To illustrate, replace 0 by a black raven, 1 by a nonblack raven, the simpler language by the general hypothesis "all ravens are black", and the more complex language by "not all ravens are black". Then the language learning problem becomes a scientific problem: a scientist is wondering whether all ravens are black, and a salient learning method for her to adopt, or at least for her to consider, is one that corresponds to enumerative induction: "Say 'yes, all ravens are black' when you have seen a sufficiently large amount of ravens and all of them are black; say 'no' when you have seen at least one counterexample." Now, note that it is possible for a (mortal or immortal) scientist to observe only black ravens (i.e. 0s) throughout her lifetime while there actually exist some nonblack ravens (i.e. 1s). Nature need not be malicious to make this happen; all it takes is just bad luck, and the bad luck could be generated, for example, by an unknown deterministic process, or by an unknown probabilistic process that involves a sequence of random variables that might or might not be i.i.d. (i.e. independent and identically distributed). But scientists are nonetheless interested in whether all ravens are black. This raises a question:

> **(Question)** Scientists seem generally interested in whether all Fs are Gs, not just whether all Fs to be actually observed in the future are Gs. So, they seem interested in some aspects of the world that might be underdetermined by the entire data stream to be presented incrementally to them (whether or not they have knowledge of the underlying mechanism that is responsible for such potential underdetermination). If so, how is the success of science possible at all?

This question has to be addressed by learning theory, if scientific inquires are learning processes and if learning theory aspires to be a general theory of successful learning. I submit that learning theory *can* address this question, and the first step to take is to examine a variety of convergence concepts and look for a new success criterion of identification in the limit. This is exactly what I want to do in this paper.

The question formulated above might invite a very simple answer, possibly from a Kantian philosopher of science:

(A Simple Kantian Answer) In order for us to think that the success of science is possible, we have to presuppose that the kind of underdetermination mentioned above does not hold in the possible world we actually live in. And, if we presuppose so, we will fall back to the easy language learning problem, making the hard problem irrelevant. Granted, this sounds like wishful thinking. But the point is that we have no alternative but to presuppose so—if we really want to think that the success of science is possible.

I used to be sympathetic to this Kantian answer until I recognized that it presupposes a lot: as a statement about the possibility of success, it tacitly presupposes that there is no more success criterion than those we have already formulated and put on the table. I propose that we challenge this tacit presupposition, look for a new success criterion, and hold on to the great tradition (1)-(3) in theoretical computer science. This is exactly what I want to do in this paper.

Here is the plan: I will first present Gold's classic definition of identification in the limit—in a way that appears complicated but is meant to reveal all the moving parts of the classic definition (section 2). Then, to find inspiration for modifying those moving parts, I will review certain concepts of limits and convergence in analysis and topology, with a focus on two very important ideas: first, the generalization from sequence convergence to net convergence; and second, the distinction between pointwise and uniform convergence (section 3). These two ideas provide a great resource of inspiration for modifying the moving parts found in Gold's classic definition, leading ultimately to my proposed definition of identification in the limit (section 4). Then I will be able to formulate the two promised theorems (section 5) and prove them (appendix A).

2 Classic Convergence to the Truth

This section presents the Gold-style learning model. The presentation might appear more complicated than necessary (just look at how long the first definition right below is). But we will gradually see its value: it reveals the moving parts of Gold's definition of identification in the limit and, thereby, suggests new definitions.

Definition 1 (Problem). A (*learning*) *problem* is a 5-tuple $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ that satisfies the following conditions (accompanied by interpretations):

- $\mathcal{H}$ is a nonempty set of *hypotheses*, understood as the theories, languages, concepts, or whatever targets to learn or choose from.

- $\mathcal{I}$ is a nonempty set of *information states*, which are possible inputs into a learning method.

- $\leq$ is a partial order over $\mathcal{I}$. Understand $i \leq i'$ to say that a learner might go from information state i to i' and, when she does, there is no loss of information.

- $\mathcal{W}$ is a nonempty set of possible states of the world, or *possible worlds* for short. Each possible world $w \in \mathcal{W}$ determines at least the correct target to learn in this world, and possibly also determines some other aspects of the world such as how information or data are presented or generated, probabilistically or deterministically or adversarially. $\mathcal{W}$ need not contain all logically possible worlds, but contains exactly those that are logically compatible with the presupposition of the problem—that is, $\mathcal{W}$ represents the content of the presupposition.

- $|\cdot|$ is an (*interpretation*) *function* defined on $\mathcal{I} \cup \mathcal{H}$, mapping all $i \in \mathcal{I}$ and all $h \in \mathcal{H}$ to subsets of $\mathcal{W}$. $|i|$ is understood to contain exactly the possible worlds in $\mathcal{W}$ that i does not rule out, or equivalently, those worlds in which the informational content of i is true. Similarly, $|h|$ is understood to contain exactly the possible worlds in $\mathcal{W}$ that makes h the correct target to learn, or equivalently, those worlds in which hypothesis h is true.

- The following four axioms hold:

  - (Nonemptiness) $|h|$ and $|i|$ are nonempty.
  - (Disjointness) $h \neq h'$ implies $|h| \cap |h'| = \varnothing$, for all $h, h' \in \mathcal{H}$. (That is, distinct hypotheses are mutually incompatible, competing with each other.)
  - (Monotonicity) $i \leq i'$ implies $|i| \supseteq |i'|$, for all $i, i' \in \mathcal{I}$. (That is, there is no loss of information in the state transition from any i to any $i' \geq i$.)
  - (Linearity) Let $\mathcal{I}(w)$ denote $\{i \in \mathcal{I} : w \in |i|\}$, the set of information states that are true at possible world w. Then each partially ordered set $(\mathcal{I}(w), \leq)$ is linear and not longer than the sequence $\omega = \{0, 1, 2, \ldots\}$ of natural numbers. (This linear order is basically the data stream to be presented in possible world w.)

Remark. All results in this paper are actually proved in a more general setting, which weakens the axiom of linearity to an axiom called *directedness*.[1] But the above definition

[1] (Directedness) Each partially ordered set $(\mathcal{I}(w), \leq)$ is directed, namely $i, j \in \mathcal{I}_w$ implies $i, j \leq k$ for some $k \in \mathcal{I}_w$.

is already general enough to capture all Gold-style language learning problems; for your reference, I present in appendix B how this can be done for all problems of passive language learning from positive examples. For the present purposes we need to focus on the easy and hard language learning problems that we informally discussed in the introduction—or their isomorphic counterparts for a raven scientist.

Definition 2 (Learning Method). A (*learning*) *method* for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ is a function $M : \mathcal{I} \to \mathcal{H} \cup \{?\}$, which is interpreted as follows:[2]

$M(i) = h$ means that M chooses hypothesis h given information/input i;
$M(i) = ?$ means that M suspends judgment given information/input i.

Definition 3 (Classic Identification in the Limit). A method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ is said to *classically converge to the truth*—or *classically identify the correct target in the limit*—just in case:

(1) for each hypothesis $h \in \mathcal{H}$,
(2) for each possible world $w \in |h|$,
(3) there exists an information state $i \in \mathcal{I}(w)$ such that
(4) $M(i') = h$ for all $i' \geq i$ in $\mathcal{I}(w)$.

If a problem admits of such a learning method, it is said to be *classically solvable*—or *learnable—in the limit*.[3]

Remark. I state this classic definition in a way meant to reveal its moving parts: the ordering of (1)-(4), and the respective roles played by the five components in a learning problem $\big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$. In section 4, I will exchange the order of (2) and (3) to generate a kind of "semi-uniform" convergence, and then modify here and there to result in the new success criterion promised above. For now, let us have a look at a toy example: leave the ordering intact and slightly modify (4) as follows, which results in a *higher* success criterion:

[2]For the problems and learning methods defined here to be interesting in computer science, we need to require that the hypotheses in $\mathcal{H}$ and the information states in $\mathcal{I}$ can in principle be encoded by natural numbers. But the results in this paper hold generally whether or not we add this requirement. Furthermore, this definition can be generalized by allowing a learning method to output not just hypotheses in $\mathcal{H}$ but also their Boolean combinations.

[3]Definitions 1-3 are essentially the order-theoretic counterparts of the topologically formulated definitions proposed in Kelly et al. (2016), provided that we include the possible generalizations mentioned so far (i.e. allowing $(\mathcal{I}(w), \leq)$ to be directed and allowing a learning method to output Boolean combinations of hypotheses)

Definition 4 (Identification with Decision). A method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ is said to *identify the correct target with decision* just in case:

> for each hypothesis $h \in \mathcal{H}$,
> for each possible world $w \in |h|$,
> there exists an information state $i \in \mathcal{I}(w)$ such that
> $M(i') = h$ for all $i' \geq i$ in $\mathcal{I}$.

(Note that M in effect *halts* at information state i, because no further information would change its mind.) If a problem admits of such a learning method, it is called *decidable*.

Remark. Many interesting problems cannot be solved by meeting the higher success criterion, identification with decision, but can only be solved by meeting the lower one, identification in the limit. The easy language learning problem is one such, as mentioned above and to be define below. Well, I will actually define an isomorphic version of it—for a raven scientist.

Imagine a scientist who wonders whether all ravens are black, and is going to observe a raven at a time. 0 represents an event of observing a black one; 1, a nonblack raven. An information state is a binary sequence of finite length. For example, $(0, 0, 1)$ denotes the information state at which the scientist has observed 2 black ravens followed by a nonblack one. Let $s = (s_1, s_2, \ldots)$ be a binary sequence of infinite length. In a possible world denoted by $(\texttt{yes}, s)$, all ravens are black and the scientist receives data in the ordering of $s_1, s_2, \ldots$, one at a time—so s only contains 0. Similarly, in a possible world denoted by $(\texttt{no}, s)$, not all ravens are black and the scientist receives data in the ordering of $s_1, s_2, \ldots$, but this time s can be any infinite binary sequence. For example, let $(0, 0, \ldots, 0, \ldots)$ denote the infinite sequence of 0s. Then $(\texttt{no}, {}_{(0,0,\ldots,0,\ldots)})$ is the unfortunate world in which not all ravens are black but the scientist will never observe a counterexample. Let $\mathcal{W}_{\text{all}}$ denote the set of all possible worlds just defined, or formally:

$$\mathcal{W}_{\text{all}} \;=\; \big\{(\texttt{yes}, {}_{(0,0,\ldots,0,\ldots)})\big\} \cup \big\{(\texttt{no}, s) : s \in \{0,1\}^{\omega}\big\}\,.$$

Example 5 (Easy Raven Problem). The easy raven problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ is defined by:

- $\mathcal{H} = \{\texttt{yes}, \texttt{no}\}$.
- $\mathcal{I}$ = the set of all binary sequences of finite length.
- $i \leq i'$ iff i is extended by i', for all sequences $i, i' \in \mathcal{I}$.

- $\mathcal{W} = \mathcal{W}_{\text{all}} \setminus \{(\texttt{no}, {}_{(0,0,\ldots,0,\ldots)})\}$, which excludes the unfortunate possible world $(\texttt{no}, {}_{(0,0,\ldots,0,\ldots)})$

  So this problem presupposes that we do not live in the unhappy world, and hence that each infinite binary sequence that might be presented to the scientist uniquely determines the true hypothesis, i.e. the correct target.

  If this problem is interpreted as the easy language learning problem, in which `yes` denotes the simple language that contains exactly one string 0 and `no` denotes the more complex language that contains exactly two strings 0 and 1, then $\mathcal{W}$ represents the presupposition that the learner is to be presented with a sound and *complete* sequence of strings in the correct language.

- $|i|$ = the set of possible worlds (h, s) in $\mathcal{W}$ such that s extends i.
  $|\texttt{yes}|$ = the set of possible worlds (h, s) in $\mathcal{W}$ such that $h = \texttt{yes}$.
  $|\texttt{no}|$ = the set of possible worlds (h, s) in $\mathcal{W}$ such that $h = \texttt{no}$.

Example 6 (Hard Raven Problem). Isomorphic to the hard language learning problem (as discussed in the introduction) is the hard raven problem, which is the same as the easy problem except that:

- $\mathcal{W} = \mathcal{W}_{\text{all}}$, which does contain the unfortunate possible world $(\texttt{no}, {}_{(0,0,\ldots,0,\ldots)})$.

Then we have the following result:

Proposition 7. *The easy raven problem is classically learnable in the limit, but the hard one is not.*

So, given this result and the discussion in the introduction, we should look for a new success criterion that complements Gold's classic one.

3 Convergence Concepts in Analysis and Topology

As mentioned in the introduction, the classic definition of identification in the limit only employs the most familiar concept of convergence—pointwise convergence of sequences—while in analysis and topology mathematicians have already worked out more elaborated concepts of convergence. This section provides a quick review of those concepts, which will become a great resource of inspiration for modifying the moving parts of the classic definition.

Definition 8 (Sequence Convergence). Let $f : \omega \to Y$ be a sequence of points in a topological space Y. Say that $f(n)$ *converges* to y as n travels upward in $(\omega, \leq)$ just in case:

> for any open neighborhood U of y in Y,
> there exists $n \in \omega$ such that
> $f(m) \in U$ for every $m \geq n$ in ω.

In the special case that Y is equipped with the discrete topology (which contains all subsets of Y as open sets), the above definition reduces to the following condition:

> there exists $n \in \omega$ such that
> $f(m) = y$ for every $m \geq n$ in ω.

Remark. This special case, in which Y is equipped with the discrete topology, is all we (and Gold) really care about here. All definitions below are formulated for this special case unless stated otherwise.

The next step is to go from convergence of a sequence to convergence of a "generalized" sequence", a.k.a. "net". A sequence is defined on a very special set of indices, namely ω, linearly ordered by $\leq$. Let us have a more general set I of indices with a weaker ordering structure $\leq$:

Definition 9 (Directed Poset). A *poset* $(I, \leq)$ consists of a set I partially ordered by $\leq$. It is *directed* just in case $i, j \in I$ implies $i, j \leq k$ for some $k \in I$.

Definition 10 (Generalized Sequence, or Net). A *generalized sequence*, or *net* $f : (I, \leq) \to Y$, is a function from a nonempty directed poset $(I, \leq)$ to a topological space Y.

Definition 11 (Net Convergence). Let $f : (I, \leq) \to Y$ be a net. Say that $f(i)$ *converges* to y as i travels in $(I, \leq)$ just in case:

> there exists i in I such that
> $f(i') = y$ for every $i' \geq i$ in I.

Remark. Note that the definition of net convergence is formally identical to that of sequence convergence, except that the underlying spaces of indices are generalized. Net convergence has a number of equivalent formulations, some of which are very sophisticated (and stated in terms of, for example, filters).[4] Let me introduce two very simple but illuminating ones.

[4]See Kelley (1991) for a review.

Definition 12 (Upper & Principal Subset). A subset S of a poset $(I, \leq)$ is *upper* just in case S is upward closed in the sense that, if $i \in S$ and $i \leq i' \in I$, then $i' \in S$. It is called a *principal* upper subset if, furthermore, it takes the following "principal" form: $S = \{i' \in I : i \leq i'\}$ for some $i \in I$.

Proposition 13 (Net Convergence Redefined by "Principal Upper"). *Let $f : (I, \leq) \to Y$ be a net. Then net convergence of f can be equivalently redefined in terms of principal upper subsets as follows. $f(i)$ converges to y as i travels in $(I, \leq)$ if and only if:*

> *there exists a principal upper subset S of $(I, \leq)$ (as a concept of "sufficiently high") such that $f(i) = y$ for every $i \in S$ (i.e. for every i that counts as "sufficiently high" in the ordering).*

Remark. The above result follows immediately from definitions. But it illustrates the idea of limits and convergence very intuitively: what we converge to is what we get when we climb up in the index set to any sufficiently high extent. The above result says that a convergence zone—or a region of sufficient height—must have a certain "shape", namely the "shape" of principal upper subsets, like the shape of ice cream cones. So we can try out different "shapes" of convergence zones. We might as well require that a convergence zone should be "dense" in the following sense:

Definition 14 (Dense Subset). A subset S of a poset $(I, \leq)$ is *dense* just in case, for each $i \in I$, there exists $i' \in S$ with $i \leq i'$.

Remark. It turns out that, if we put the concept of net convergence on an operating table and change convergence zones from principal upper subsets to dense upper subsets, then it remains to be the same convergence concept, as reported below.

Proposition 15 (Net Convergence Redefined by "Dense Upper"). *Let $f : (I, \leq) \to Y$ be a net. Then net convergence of f can be equivalently redefined in terms of dense upper subsets as follows. $f(i)$ converges to y as i travels in $(I, \leq)$ if and only if:*

> *there exists a dense upper subset S of $(I, \leq)$ such that*
> *$f(i) = y$ for every $i \in S$.*

Remark. The proof is routine and elementary (but it makes essential use of the assumption that $(I, \leq)$ is directed).

We have talked about distinct *zones* of convergence. Now turn to two *flavors* of convergence, pointwise and uniform:

Definition 16 (Pointwise vs. Uniform Convergence). Let f_i and g be functions from X to Y, where i is in a directed poset $(I, \leq)$ and, hence, the mapping from i to f_i is a net. Say that $(f_i)_{i \in I}$ *converges to* g *pointwisely* if it satisfies the first condition below; say that it does so *uniformly* if it satisfies the second condition below:

A. (Pointwise Convergence)
For all $x \in X$,
  there exists $i \in I$ such that
    $f_{i'}(x) = g(x)$ for every $i' \geq i$ in I.

B. (Uniform Convergence)
There exists $i \in I$ such that,
  for all $x \in X$,
    $f_{i'}(x) = g(x)$ for every $i' \geq i$ in I.

Remark. Note that the formal difference between pointwise convergence and uniform convergence lies in the order of quantifiers 'for all' and 'there exists'. Pointwise convergence requires that for each $x \in X$ there be some convergence zone—some concept of "sufficiently high"—that takes care of x itself, while uniform convergence requires that there be a single one that takes care of all $x \in X$.

4 Toward a New Convergence Concept in Learning Theory

I have presented the *ideas* behind the formal definitions of limits and convergence in mathematics. So we are ready to get back to learning theory. With those ideas, it is not hard to see that Gold's classic definition of identification in the limit is only a particular implementation of the following, informal template:

A. (Pointwise Convergence to the Truth)

No matter which hypothesis $h \in \mathcal{H}$ is true,
  no matter which world $w \in |h|$ is actual,
    there exists a concept *SI* of "sufficient information" associated with w such that,
      M outputs the truth h
      as long as the input i is "sufficiently informative" and true at w.

In this template, each world—or point—is associated with a sufficiency concept. So this deserves to be called "pointwise". Once this is seen, it is only natural to seek a "more uniform" version of it, exchanging the second line (which is a universal quantification over w) and the third line (which is an existential quantification over SI). Then we have:

B. (Semi-Uniform Convergence to the Truth)
No matter which hypothesis $h \in \mathcal{H}$ is true,
there exists a concept SI of "sufficient information" associated with h such that,
no matter which world $w \in |h|$ is actual,
M outputs the truth h
as long as the input i is "sufficiently informative" and true at w.

The goal of this section is to implement these two templates in a way that "works" as promised in the introductory section.

Remark. With these two informal templates, the next obvious move is to look for a workable definition of concepts of "sufficient information" or "convergence zone". But we need some preliminaries. The third line of template A involves a concept of sufficient information associated with a world w, so this concept is a subset of $\mathcal{I}(w) = \{i \in \mathcal{I} : w \in |i|\}$, the set of information states that are true at possible world w. Now, the second line of template B involves a concept of sufficient information associated with, not a world, but a hypothesis h, so this concept should be a subset of $\mathcal{I}(h)$, which is defined as the set of information states that might be true given hypothesis h, namely:

$$\mathcal{I}(h) \;=\; \bigcup_{w \in |h|} \mathcal{I}(w)\,.$$

Note that, although $(\mathcal{I}(w), \leq)$ is assumed to be linear, $(\mathcal{I}(h), \leq)$ could be a mere poset, neither linear nor directed. Then the above two informal templates A and B can be made "almost formal" as follows, pending the two blanks to be filled in.

Definition Template 17. The following conditions are two templates for defining *convergence to the truth* or *identification in the limit*:

A. (Pointwise Convergence to the Truth)

For each hypothesis $h \in \mathcal{H}$,
for each possible world $w \in |h|$,
there exists a _____ upper subset SI of linear poset $(\mathcal{I}(w), \leq)$ such that
$M(i) = h$ for each information state $i \in SI$ (that is true at w).

B. (Semi-Uniform Convergence to the Truth)

For each hypothesis $h \in \mathcal{H}$,
  there exists a _____ upper subset SI of poset $(\mathcal{I}(h), \leq)$ such that,
    (for each possible world $w \in |h|$,)
      $M(i) = h$ for each information state $i \in SI$ (that is true at w).

Remark. The parts enclosed in parentheses are redundant, which can be easily proved from the definitions of problems and $\mathcal{I}(\cdot)$. But it is important to keep them here—to emphasize the parallel to the informal templates. Note that the term 'upper' has been built into the templates, and rightly so. For if a state i is more informative than some state with "sufficient information" then i itself should also be taken as a state with "sufficient information".

The question is how to fill in the blanks.

Template A seems to me useless. If we fill in its blank with 'principal' or 'dense', then there will be nothing new—we will obtain something equivalent to Gold's classic definition because $(\mathcal{I}(w), \leq)$ is assumed to be linear.[5] Might template A be made useful by some order-theoretic concept other than 'principal' and 'dense'? I do not see any.

For template B, we do *not* want to fill in the blank with 'principal'. Here is why: If a convergence zone in a *partially ordered* set $(\mathcal{I}(w), \leq)$ is allowed to be a principal upper subset, then in effect a convergence zone is allowed to be a *narrow* ice cream cone in an ambient, possibly *wide* poset $(\mathcal{I}(w), \leq)$. This would make the convergence criterion very weak, too weak to result in what deserves to be called a success criterion.

So I propose that, for template B, we fill in the blank with 'dense', whose definition (as presented above) requires that a convergence zone be wide, almost as wide as the ambient poset. In fact, the concept of 'dense' also serves to capture another intuitive idea about sufficient information that *complements* the intuitive idea captured by 'upper'. The idea is that, *whatever information we have, in principle it should be always possible for us to obtain more information and have sufficient information.* Or more formally, whenever we are at some information state i in $\mathcal{I}(h)$, it should be always possible for us to obtain more information and go into some information state $i' \geq i$ in $\mathcal{I}(h)$ such that i' contains sufficient information. Namely, a concept of sufficient information, as a subset SI of $(\mathcal{I}(h), \leq)$, should be dense. So I propose the following definition:

Definition 18 (Semi-Uniform Convergence to the Truth). A method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ is said to *semi-uniformly converge to the truth*—or *semi-*

[5] This claim still holds even if $(\mathcal{I}(w), \leq)$ is only assumed to be directed, thanks to proposition 15.

uniformly identify the correct target in the limit—just in case it satisfies the following implementation of template B:

> For each hypothesis $h \in \mathcal{H}$,
> there exists a dense upper subset SI of $(\mathcal{I}(h), \leq)$ such that,
> (for each possible world $w \in H$,)
> $M(i) = h$ for each information state $i \in SI$ (that is true at w).

If a problem admits of such a learning method, it is called *solvable*—or *learnable*—*in the limit semi-uniformly.*

Then we have:

Proposition 19 (Hard Raven Problem Solved Semi-Uniformly). *The hard raven problem is solvable in the limit semi-uniformly.*

This result is witness by any learning method of the following form: "Say `yes` when you have seen n ravens and all of them are black; say `no` when you have seen at least one nonblack raven", where n is a natural number.

5 Toward a New Success Criterion

So the hard raven problem can be solved by meeting the new convergence criterion defined above. It remains to deliver the last part of the promise: to provide reasons for thinking that the new convergence criterion is high enough—demanding enough—to deserve to be called a *success* criterion. This section presents the reasons I have so far, culminating in the two theorems of this paper.

The first result says that between the two criteria of identification in the limit (the semi-uniform and the classic), neither is strictly more demanding than the other. The following example will be used to prove this claim.

Example 20 (Even-vs-Odd Raven Problem). Now the scientist wonders whether the number of nonblack ravens in the world is even or odd, while presupposing that this number is finite (and thus either even or odd), and presupposing that data presentations are complete (as in the easy raven problem). The even-vs-odd raven problem is defined as follows:

- $\mathcal{H} = \{\texttt{even}, \texttt{odd}\}$.

- $\mathcal{I}$ = the set of all binary sequences of finite length (as usual).
- $i \leq i'$ iff i is extended by i', for all sequences $i, i' \in \mathcal{I}$ (as usual).
- $\mathcal{W} = \{s \in \{0,1\}^\omega : \text{the occurrences of 1 in } s \text{ are finite in number}\}$
- $|i|$ = the set of possible worlds s in $\mathcal{W}$ such that s extends i.
  $|\texttt{even}|$ = the set of s in $\mathcal{W}$ such that the occurrences of 1 in s are even in number.
  $|\texttt{odd}|$ = the set of s in $\mathcal{W}$ such that the occurrences of 1 in s are odd in number.

Theorem 21 (Independence Result). *Semi-uniform learnability in the limit and its classical counterpart are independent of each other; that is, neither implies the other. In particular, consider the three raven problems defined above: the hard, the easy, and the even-vs-odd. Their respective learnability in the limit is summarized by the following table:*

Raven Problems:	*Hard*	*Easy*	*Even-vs-Odd*
Classically Learnable in the limit?	*No*	***Yes***	***Yes***
Semi-Uniformly Learnable in the limit?	***Yes***	***Yes***	*No*

Remark. As the proof of this proposition suggests (see appendix A), to solve a problem in the limit semi-uniformly, a learning method is required to work with "somewhat uniform" concepts of sufficient information, or zones of convergence, or bounds on convergence. This turns out to be a very demanding requirement when the scientist is tackling the even-vs-odd raven problem: there has to be a single "bound" that works for all information states compatible with `even`; and similarly, there has to be another "bound" that works for all information states compatible with `odd`.

The second result concerns a kind of learning process that cognitive scientists and learning theorists call *U-shaped learning* (Carlucci et al. 2005, Carlucci et al. 2013). It consists of conjecturing a hypothesis, then retracting it, and then conjecturing it again. It can be defined formally as follows.

Definition 22 (U-Shaped Learning). Let M be a learning method for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$. An *instance of U-shaped learning* of M is a sequence (i_1, i_2, i_3) of three information states in $\mathcal{I}$ such that:

1. $i_1 \leq i_2 \leq i_3$,
2. $M(i_1) \neq M(i_2) \neq M(i_3)$,
3. $M(i_1) = M(i_3) \in \mathcal{H}$.

Remark. Following Kelly et al. (2016), I submit that such a learning process is undesirable and, other things being equal, it had better be avoided or removed if possible. Removal of U-shaped learning is defined as follows.

Definition 23 (Removal of U-shaped Learning). For any learning methods M and M^*, say that M^* can be constructed from M by *removing U-shaped learning* just in case:

1. $M(i) = \texttt{?}$ implies $M^*(i) = \texttt{?}$,
2. $M(i) = h$ implies $M^*(i) = h$ or $\texttt{?}$,
3. M^* has no instance of U-shaped learning.

Then we have:

Theorem 24 (Persistence under Removal of U-Shaped Learning). *A method M solves a problem $\mathcal{P}$ in the limit semi-uniformly only if we can construct a method M^* from M by removing U-shaped learning such that M^* also solves $\mathcal{P}$ in the limit semi-uniformly.*

Remark. So semi-uniform identification in the limit seems to set a standard for problem solving that is by no means low: if a learning method achieves this standard, it has to do so *without* involving U-shaped learning in any essential, unremovable way. We may say that semi-uniform identification in the limit has the property called *persistence under removal of U-shaped learning*. By way of contrast, this property is not shared by classic identification in the limit. Here is a proof sketch: U-shaped learning is involved in *all* learning methods that solve the even-vs-odd raven problem in the limit classically. So, when tackling this problem, removal of U-shaped learning implies no longer having a method that solves it in the limit classically.

So, thanks to the above two theorems, semi-uniform identification in the limit is by no means a low criterion. I submit that it is high enough to deserve to be called a success criterion for problem solving. To clarify: I am *not* saying that we should always be satisfied with a solution to a problem that meets only the new success criterion. If a problem can be solved by meeting semi-uniform identification in the limit because it can be solved by meeting a strictly higher criterion, such as identification with decision, then we ought to strive for meeting the higher one. A success criterion marks an achievement, but does not necessarily mark an achievement that we should be satisfied with. We ought to strive for meeting the highest achievable success criterion—or a highest one, if success criteria are partially ordered.

I also submit that the two kinds of identification in the limit, the classic and the semi-uniform, are *complementary* criteria of success. They set very similar standards of success, but nonetheless each describes a unique way of success—so unique that meeting one does not entail meeting the other. *Meeting any one of these two is good; meeting both at the same time, if possible, is even better.* This is the sense in which these two success criteria are complementary.

That said, it might be possible for their relationship to turn competitive occasionally. Imagine that we have a problem and there are learning methods that solve it in the limit—some do it classically, some do it semi-uniformly, but no single one can do it both classically and semi-uniformly. This raises two questions. *Does there exist such a problem? If yes, which of the two success criteria should be met at the cost of the other?* I do not have an answer at the moment.

6 Guideline for Exploration

At this point you might be wondering whether there is any other success criterion that corresponds to yet another sense of identification in the limit. There might be. But what is important is that now we have a guideline for exploration:

1. Try an ordering of the four quantifiers involved in the classic definition. Use the ordering to create a definition template like those in definition 25, with a blank to be filled in.

2. Fill in the blank with an order-theoretic term (such as 'principal', 'dense', and the like) and obtain a new definition.

3. With the new definition, try to answer the following questions:

   - Does the new definition correspond to a genuinely new concept of identification in the limit, inequivalent to any old one we have already had?
   - Does the new definition sets a standard that is high enough to deserve to be called a success criterion for problem solving?
   - Does there exist an interesting problem that can be solved by meeting the new criterion but cannot be solved by meeting at least one of the old success criteria?

4. If the answers are all positive, we have obtained a new success criterion that ought to be included in learning theory.

For example, step 1 can result in not just the pointwise template A and the semi-uniform template B, but also the following, uniform template:

Definition Template 25. The following condition is a(nother) template for defining *convergence to the truth* or *identification in the limit*:

C. (Uniform Convergence to the Truth)

There exists a _____ upper subset SI of poset $(\mathcal{I}, \leq)$ such that,
for each hypothesis $h \in \mathcal{H}$,
for each possible world $w \in |h|$,
$M(i) = h$ for each information state $i \in SI$ that is true at w.

I have not been able to turn this template into a definition that yields positive answers to all the three questions at step 3. It might be because I have not tried hard enough, or it might not.

7 Closing: One Dense Upper Ring to Rule Them All?

The title of this paper, together with the parentheses therein, suggests that I will provide a very general definition of limits and convergence *per se*. That is right. I propose the following definition and, from now on, relax the restriction to discrete topology.

Definition 26 (Generalized Convergence and Limit). Let f be a function from a nonempty poset $(I, \leq)$ to an arbitrary topological space Y. Say that $f(i)$ *converges* to y as i travels in $(I, \leq)$ (in the sense of *generalized convergence*) just in case:

for any open neighborhood U of y,
there exists a dense upper subset S of $(I, \leq)$ such that
$f(i) \in U$ for all $i \in S$.

If such a y exists uniquely, also say that y is the *generalized limit* of $f(i)$ as i travels in $(I, \leq)$.

The above is the concept of convergence that underlines the definition of semi-uniform identification in the limit: the index i travels in a *partially ordered set* I in general, a convergence zone is required to be dense and upper, and the underlying topology of the codomain is discrete.

When the domain $(I, \leq)$ of f is not just partially ordered but directed, generalized convergence reduces to net convergence. To be more specific, let me state the full definition of net convergence without the restriction to discrete topology:

Definition 27 (Net Convergence). Let $f : (I, \leq) \to Y$ be a net, where Y is an arbitrary topological space. Say that $f(i)$ *converges* to y as i travels in $(I, \leq)$ (in the sense of *net convergence*) just in case:

> for any open neighborhood U of y in Y,
> there exists i in I such that
> $f(i') \in U$ for every $i' \geq i$ in I.

Then we have the reduction result I promised above:

Proposition 28 (Net Convergence Reduced by Generalized Convergence). *Let f be a function from a nonempty poset $(I, \leq)$ to an arbitrary topological space Y. Suppose that f is a net, i.e. its domain $(I, \leq)$ is directed. Then $f(i)$ converges to y as i travels in $(I, \leq)$ in the sense of net convergence if, and only if, it does so in the sense of generalized convergence.*

This result generalizes proposition 15 precisely by relaxing the restriction to discrete topology.

It follows that in principle we can use generalized convergence to do whatever we can do with net convergence, which is by far the most general concept of convergence used in analysis and topology.[6] For example, net convergence can be used to define continuous functions between arbitrary topological spaces, and it provides the standard definition of the Riemann integral. When the domain $(I, \leq)$ of f is not just directed but an ω-sequence, net convergence reduces to sequence convergence. See appendix C for the details of the examples just mentioned.

So the concept of generalized convergence is very general, subsuming concepts of convergence in analysis and topology and, I hope, in learning theory.

References

Carlucci, L. and Case, J. (2013) "On the Necessity of U-Shaped Learning", *Topics in Cognitive Science*, 5(1): 56-88.

[6] This is *not* a claim about which one is more convenient to use in analysis and topology.

Carlucci, L., Case, J., Jain, S., and Stephan, F. (2005) "Non U-Shaped Vacillatory and Team Learning", *Algorithmic Learning Theory*, 241-255, Springer Berlin Heidelberg.

Gold, E. M. (1967) "Language identification in the limit", *Information and Control*, 10(5): 447-474.

Kelley, J. L. (1991) *General Topology*, Springer.

Kelly, T. K, Genin, K. and Lin, H. (2016) "Realism, Rhetoric, and Reliability", *Synthese* 193 (4):1191-1223.

Valiant, L. (1984) "A Theory of the Learnable", *Communications of the ACM* 27(11): 1134-1142.

A Proofs of Results

Proof of Proposition 7. The claim about the easy raven problem is well known. The non-learnability result of the hard raven problem follows from the fact that any method M is doomed to have this property: $M(i)$ converges to the truth as i travels in the data stream in possible world $(\texttt{yes}, {}_{(0,0,\ldots,0,\ldots)})$ iff it fails to do so in possible world $(\texttt{no}, {}_{(0,0,\ldots,0,\ldots)})$. □

Proof of Proposition 15. This proposition is a special case of proposition 28—the special case in which codomain Y is a topological space equipped with the discrete topology. See below for the proof of proposition 28, which is fully general and applicable to all topological spaces. □

Proof of Proposition 19. Consider this learning method: "Say `yes` when you have not seen any nonblack raven; otherwise say `no`." This method converges to the truth semi-uniformly for the following two reasons. First, concerning output `yes`, this method has a "convergence zone" $SI = \mathcal{I}(\texttt{yes})$, which is (trivially) a dense upper subset of $(\mathcal{I}(\texttt{yes}), \leq)$. Second, concerning output `no`, this method has a "convergence zone" $SI' = \{i \in \mathcal{I} : i \text{ extends } 0^n1 \text{ for some } n \in \omega\}$, which is a dense upper subset of $(\mathcal{I}(\texttt{no}), \leq)$. (This is because $\mathcal{I}(\texttt{no}) = \mathcal{I}$, the set of all finite binary sequences.) □

Proof of Theorem 21. Thanks to proposition 19, it suffices to prove the part for the even-vs-odd raven problem. Suppose, for *reductio*, that some learning method M for the even-vs-odd raven problem solves it in the limit semi-uniformly. So there exists a dense upper subset SI of $(\mathcal{I}(\texttt{even}), \leq)$ such that $M(i) = \texttt{even}$ for all $i \in SI$. Similarly, there exists a dense upper subset SI of $(\mathcal{I}(\texttt{odd}), \leq)$ such that $M(i) = \texttt{odd}$ for all $i \in SI$. But

here is the key: $\mathcal{I}(\texttt{even}) = \mathcal{I}(\texttt{odd}) = \mathcal{I}$. So, by the order-theoretic result that two dense upper subsets SI, SI' of a nonempty poset $(\mathcal{I}, \leq)$ share at least one common element i, we have that $M(i) = \texttt{even} = \texttt{odd}$, contradiction. The classical learnability of the even-vs-odd problem is witnessed by this learning method: "Say `even` when you have seen an even number of nonblack ravens; otherwise say `odd`." □

Proof of Theorem 24. Suppose that method M identifies the correct target for problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ in the limit semi-uniformly. Then, for each hypothesis $h \in \mathcal{H}$, $(\mathcal{I}(h), \leq)$ has a dense upper subset SI_h such that, for all $i \in SI_h$, $M(i) = h$. Construct a learning method M^* from M as follows:

$$M^*(i) \;=\; \begin{cases} h & \text{if } i \in SI_h \\ \texttt{?} & \text{otherwise.} \end{cases}$$

By construction, M^* solves the same problem in the limit semi-uniformly. And, by construction again, it satisfies the first two of the three conditions that jointly define "constructability by removing U-shaped learning". So it suffices to show that M^* satisfies the last of the three conditions, namely that M^* has no instance of U-shaped learning.

Suppose, for *reductio*, that M^* has an instance (i_1, i_2, i_3) of U-shaped learning. Then there exists $h^* \in \mathcal{H}$ such that $M^*(i_1) = M^*(i_3) = h^*$. So, by construction, both i_1 and i_3 are in SI_{h^*}. It follows that $i_3 \in \mathcal{I}(h^*)$. So, by the definition of $\mathcal{I}(\cdot)$, there exists $w^* \in \mathcal{W}$ such that $w^* \in |i_3| \cap |h^*|$. Furthermore, since (i_1, i_2, i_3) is an instance of U-shaped learning of M^*, we have $i_1 \leq i_2 \leq i_3$. So, by Monotonicity, we have $|i_1| \supseteq |i_2| \supseteq |i_3|$. It follows that $|i_1| \supseteq |i_2| \supseteq |i_3| \ni w^* \in |h^*|$. So $w^* \in |i_2| \cap |h^*|$, and hence $i_2 \in \mathcal{I}(h^*)$. To summarize what we have obtained so far: SI_{h^*} contains i_1 and is an upper subset of $(\mathcal{I}(h^*), \leq)$, and $i_1 \leq i_2 \in \mathcal{I}(h^*)$. It follows from upward closure that SI_{h^*} contains i_2, too. So, by construction, $M^*(i_1) = M^*(i_2) = M^*(i_3) = h^*$. But this contradicts the *reductio* hypothesis that (i_1, i_2, i_3) is an instance of U-shaped learning of M^*. □

Proof of Proposition 28. The ($\Rightarrow$) side follows immediately from the order-theoretic result that every principal upper subset of a nonempty directed poset is dense. The ($\Leftarrow$) side follows immediately from the order-theoretic result that every dense upper subset of a nonempty directed poset is nonempty (because of denseness) and thus (by upward closure) includes a principal upper subset. □

B Problems of Passive Language Learning from Positive Examples

Suppose that we are given a fixed, finite set of alphabets. Strings are finite sequences of alphabets. A language is a set of strings. A problem of *passive language learning from positive examples* is a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{I}, \leq, \mathcal{W}, |\cdot|\big)$ that takes the following form:

- $\mathcal{H} = \{L_0, L_1, \ldots\}$, a set of languages.
- $\mathcal{I}$ = the set of all finite sequences of strings. An element $i = (s_1, \ldots, s_n)$ of $\mathcal{I}$ represents the information state at which the strings in i have been presented to the learner in this order: $s_1, \ldots$, and then s_n.
- $i \leq i'$ iff i is extended by i'.
- W is the set of all possible worlds of this form: (L_n, s), where L_n is a language and $s = (s_1, s_2, \ldots)$ is an infinite sequence of strings such that $\{s_1, s_2, \ldots\} = L_n$. In a possible world (L_n, s), the correct language to learn is L_n and the learner is presented with the strings in L_n, one at a time, according to the ordering in s. The learner does not know which possible world is the actual one. The identity requirement that $\{s_1, s_2, \ldots\} = L_n$ amounts to the assumption of soundness and completeness discussed in the introductory section.
- $|i|$ is the set of all possible worlds (L_k, s) such that s extends i.
  $|L_n|$ is the set of all possible worlds (L_k, s) such that $L_k = L_n$.

C Examples of What Can be Defined by Generalized Limits

- A method M identifies the correct target in the limit semi-uniformly if, and only if, every hypothesis $h \in \mathcal{H}$ is the generalized limit of $M(i)$ as i travels in poset $(\mathcal{I}(h), \leq)$.
- A method M identifies the correct target in the limit classically if, and only if, for any hypothesis $h \in \mathcal{H}$ and any possible world $w \in |h|$, h is the generalized limit of $M(i)$ as i travels in directed poset $(\mathcal{I}(w), \leq)$.

- $\lim_{x\to\infty} f(x)$ is the generalized limit of $f(x)$ as x travels in linear poset $(\mathbb{R}, \leq)$, where $\leq$ is the natural ordering on $\mathbb{R}$.
- $\lim_{x\to a} f(x)$ is the generalized limit of $f(x)$ as x travels in directed poset $(\mathbb{R}\setminus\{a\}, \leq_a)$, where the ordering $\leq_a$ is defined as follows: $x > x'$ iff $|x-a| < |x'-a|$.
- Riemann integral $\int_a^b f(x)\,dx$ is the generalized limit of the Riemann sum of f over tagged partition π, as π travels in directed poset $\big(\Pi[a,b], \sqsubseteq\big)$, where $\Pi[a,b]$ is the set of tagged partitions over closed interval $[a,b]$, which is partially and directedly ordered by the refinement relation $\sqsubseteq$ between tagged partitions.