Mind Change Efficient Learning

Wei Luo and Oliver Schulte

School of Computing Science, Simon Fraser University,
Vancouver, Canada
`wluoa,oschulte@cs.sfu.ca`

Abstract. This paper studies efficient learning with respect to mind changes. Our starting point is the idea that a learner that is efficient with respect to mind changes minimizes mind changes not only globally in the entire learning problem, but also locally in subproblems after receiving some evidence. Formalizing this idea leads to the notion of *uniform mind change optimality*. We characterize the structure of language classes that can be identified with at most α mind changes by some learner (not necessarily effective): A language class $\mathcal{L}$ is identifiable with α mind changes iff the accumulation order of $\mathcal{L}$ is at most α. Accumulation order is a classic concept from point-set topology. To aid the construction of learning algorithms, we show that the characteristic property of uniformly mind change optimal learners is that they output conjectures (languages) with maximal accumulation order. We illustrate the theory by describing mind change optimal learners for various problems such as identifying linear subspaces and one-variable patterns.

1 Introduction

One of the goals of computational learning theory is to design learning algorithms for which we can provide performance guarantees. Identification in the limit is a central performance goal in Gold's language learning paradigm [9]. A well-studied refinement of this notion is *identification with bounded mind changes* [8, 1]. In this paper we investigate a further refinement that we term uniform mind change optimality (UMC-optimality). Briefly, a learner is UMC-optimal if the learner achieves the best possible mind change bound not only for the entire problem, but also relative to data sequences that the learner may observe.

The general theory in this paper has two main goals. (1) To provide necessary and sufficient conditions for a language collection to be identifiable with a given (ordinal) mind-change bound by some learner (not necessarily effective). (2) To provide necessary and sufficient conditions for a learner to be UMC-optimal. The results addressing (1) help us determine when a UMC-optimal learning algorithm exists, and the results addressing (2) help us to construct optimal learning algorithms when they do exist.

We situate our study in the framework of point-set topology. Previous work has shown the usefulness of topology for learning theory [25, Ch.10], [21, 14, 4]. We show how to view a language collection as a topological space; this allows us to apply Cantor's classic concept of *accumulation order* which assigns an ordinal $\mathrm{acc}(\mathcal{L})$ to a language collection, if $\mathcal{L}$ has bounded accumulation order. We show that a language collection $\mathcal{L}$ is identifiable with mind change bound α by a learner if and only if $\mathrm{acc}(\mathcal{L}) = \alpha$.

This result establishes a purely information-theoretic and structural necessary condition for identification with bounded mind changes. Based on the concept of accumulation order, we provide necessary and sufficient conditions for a learner to be UMC-optimal. These results show that UMC-optimality strongly constrains the conjectures of learners. We illustrate these results by analyzing various learning problems, such as identifying a linear subspace and a one-variable pattern.

The paper is organized as follows. Sect. 2 reviews standard concepts for language identification and presents our definition of mind change optimality. Then we establish the correspondence between mind change complexity and accumulation order. Sect. 4 gives necessary and sufficient conditions for a learner to be uniformly mind change optimal. Finally, we describe a general approach to constructing UMC-optimal effective learners and illustrate it with one-variable pattern languages.

2 Preliminaries: Language Identification

2.1 Standard Concepts

We employ notation and terminology from [12], [20, Ch.1], and [9]. We write $\mathbb{N}$ for the set of natural numbers: $\{0, 1, 2, ...\}$. The symbols $\subseteq, \supseteq, \subset, \supset$, and $\emptyset$ respectively stand for subset, superset, proper subset, proper superset, and the empty set. We view a language as a set of strings. We identify strings with natural numbers encoding them. Thus we define a **language** to be a subset of $\mathbb{N}$ and write L for a generic language [9, p.449]. A **language learning problem** is a collection of languages; we write $\mathcal{L}$ for a generic collection of languages. A **text** T is a mapping of $\mathbb{N}$ into $\mathbb{N} \cup \{\#\}$, where $\#$ is a symbol not in $\mathbb{N}$. (The symbol $\#$ models pauses in data presentation.) We write $\mathrm{content}(T)$ for the intersection of $\mathbb{N}$ and the range of T. A text T is **for** a language L iff $L = \mathrm{content}(T)$. The initial sequence of text T of length n is denoted by $T[n]$. The set of all finite initial sequences over $\mathbb{N} \cup \{\#\}$ is denoted by SEQ. We let σ and τ range over SEQ. We write $\mathrm{content}(\sigma)$ for the intersection of $\mathbb{N}$ and the range of σ. The initial sequence of σ of length n is denoted by $\sigma[n]$.

We say that a language L is **consistent** with σ iff $\mathrm{content}(\sigma) \subseteq L$. We write $\sigma \subset T$ or $T \supset \sigma$ to denote that text T extends initial sequence σ. For a language collection $\mathcal{L}$, the set of all finite sequences consistent with $\mathcal{L}$ is denoted by $\mathrm{SEQ}(\mathcal{L})$ (i.e., $\mathrm{SEQ}(\mathcal{L}) \equiv \{\sigma \in \mathrm{SEQ} : \exists L \in \mathcal{L}. \mathrm{content}(\sigma) \subseteq L\}$).

Examples.
(1) Let $L_i \equiv \{n : n \geq i\}$, where $i \in \mathbb{N}$; we use COINIT to denote the class of languages $\{L_i : i \in \mathbb{N}\}$ [1, p.324].
(2) In the n-dimensional linear space $\mathbb{Q}^n$ over the field of rationals $\mathbb{Q}$, we can effectively encode every vector $\boldsymbol{v}$ by a natural number. Then a linear subspace of $\mathbb{Q}^n$ corresponds to a language. We write LINEAR_n for the collection of all (encodings of) linear subspaces of $\mathbb{Q}^n$.

A **learner** is a function that maps a finite sequence to a language or the question mark ?, meaning "no answer for now". We normally use the Greek letter Ψ and variants to denote a learner. Our term "learner" corresponds to the term "scientist" in [20, Ch.2.1.2]. In typical applications we have available a syntactic representation for each member of the language collection $\mathcal{L}$ under investigation. In such settings we assume the existence of an index for each member of $\mathcal{L}$, that is, a function $index : \mathcal{L} \mapsto \mathbb{N}$ (cf. [10, p.18]), and we can take a **learning function** to be a function that maps a finite sequence to an index for a language (learning functions are called "scientists" in [10, Ch.3.3]). A computable learning function is a **learning algorithm**. We use the general notion of a learner for more generality and simplicity until we consider issues of computability.

Let $\mathcal{L}$ be a collection of languages. A learner Ψ **for** $\mathcal{L}$ is a mapping of SEQ into $\mathcal{L} \cup \{?\}$. Thus the learners we consider are class-preserving; for the results in this paper, this assumption carries no loss of generality. Usually context fixes the language collection $\mathcal{L}$ for a learner Ψ.

We say that a learner Ψ **identifies** a language L on a text T for L, if $\Psi(T[n]) = L$ for all but a finite number of stages n. Next we define identification of a language collection relative to some evidence.

Definition 1. *A learner Ψ identifies $\mathcal{L}$ given $\sigma \iff$ for every language $L \in \mathcal{L}$, and for every text $T \supset \sigma$ for L, we have that Ψ identifies L on T.*

Thus a learner Ψ identifies a language collection $\mathcal{L}$ if Ψ identifies $\mathcal{L}$ given the empty sequence Λ.

Examples.

(1) The following learner Ψ_{CO} identifies COINIT: If $\mathrm{content}(\sigma) = \emptyset$, then $\Psi_{\mathrm{CO}}(\sigma)$:=?. Otherwise set $m := \min(\mathrm{content}(\sigma))$, and set $\Psi_{\mathrm{CO}}(\sigma) := L_m$.

(2) Let $\mathrm{vectors}(\sigma)$ be the set of vectors whose code numbers appear in σ. Then define $\Psi_{\mathrm{LIN}}(\sigma) = \mathrm{span}(\mathrm{vectors}(\sigma))$, where $\mathrm{span}(V)$ is the linear span of a set of vectors V. The learner Ψ_{LIN} identifies LINEAR_n. The problem of identifying a linear subspace of reactions arises in particle physics, where it corresponds to the problem of finding a set of conservation principles governing observed particle reactions [17, 27]. Interestingly, it appears that the theories accepted by the particle physics community match the output of Ψ_{LIN} [28, 26].

A learner Ψ **changes its mind** at some nonempty finite sequence $\sigma \in \mathrm{SEQ}$ if $\Psi(\sigma) \neq \Psi(\sigma^-)$ and $\Psi(\sigma^-) \neq ?$, where σ^- is the initial segment of σ with σ's last element removed [7, 1]. (No mind changes occur at the empty sequence Λ.)

Definition 2 (based on [1]). *Let Ψ be a learner and c be a function that assigns an ordinal to each finite sequence $\sigma \in \mathrm{SEQ}$.*

1. *c is a **mind-change counter** for Ψ and $\mathcal{L}$ if $c(\sigma) < c(\sigma^-)$ whenever Ψ changes its mind at some nonempty sequence $\sigma \in \mathrm{SEQ}(\mathcal{L})$. When $\mathcal{L}$ is fixed by context, we simply say that c is a mind change counter for Ψ.*
2. *Ψ identifies a class of languages $\mathcal{L}$ **with mind-change bound** α given $\sigma \iff \Psi$ identifies $\mathcal{L}$ given σ and there is a mind-change counter c for Ψ and $\mathcal{L}$ such that $c(\sigma) = \alpha$.*
3. *A language collection $\mathcal{L}$ is **identifiable with mind change bound** α given $\sigma \iff$ there is a learner Ψ such that Ψ identifies $\mathcal{L}$ with mind change bound α given σ.*

Examples.
(1) For COINIT, define a counter c_0 as follows: $c_0(\sigma) := \omega$ if $\mathrm{content}(\sigma) = \emptyset$, where ω is the first transfinite ordinal, and $c_0(\sigma) := \min(\mathrm{content}(\sigma))$ otherwise. Then c_0 is a mind change counter for Ψ_{CO} given Λ. Hence Ψ_{CO} identifies COINIT with mind change bound ω (cf. [1, Sect.1]).
(2) For LINEAR_n, define the counter $c_1(\sigma)$ by $c_1(\sigma) := n - \dim(\mathrm{span}(\mathrm{vectors}(\sigma)))$, where $\dim(V)$ is the dimension of a space V. Then c_1 is a mind change counter for Ψ_{LIN} given Λ, so Ψ_{LIN} identifies LINEAR_n with mind change bound n.
(3) Let FIN be the class of languages $\{D \subseteq \mathbb{N} : D \text{ is finite}\}$. Then a learner that always conjectures $\mathrm{content}(\sigma)$ identifies FIN. However, there is no mind change bound for FIN [1].

2.2 Uniform Mind Change Optimality

In this section we introduce a new identification criterion that is the focus of this paper. Our point of departure is the idea that learners that are efficient with respect to mind changes should minimize mind changes not only globally in the entire learning problem but also locally after receiving specific evidence. For example, in the COINIT problem, the best global mind change bound for the entire problem is ω [1, Sect.1], but after observing initial data $\langle 5 \rangle$, a mind change efficient learner should succeed with at most 5 more mind changes, as does Ψ_{CO}. However, there are many learners that require more than 5 mind changes after observing $\langle 5 \rangle$ yet still succeed with the optimal mind change bound of ω in the entire problem.

To formalize this motivation, consider a language collection $\mathcal{L}$. If a mind change bound exists for $\mathcal{L}$ given σ, we write $\mathrm{MC}_{\mathcal{L}}(\sigma)$ for the least ordinal α such that $\mathcal{L}$ is identifiable with α mind changes given σ. It may be natural to require that a learner should succeed with $\mathrm{MC}_{\mathcal{L}}(\sigma)$ mind changes after each data sequence $\sigma \in \mathrm{SEQ}(\mathcal{L})$; indeed the learner Ψ_{CO} achieves this performance for COINIT. However, in general this criterion appears too strong. The reason is the following possibility: A learner Ψ may output a conjecture $\Psi(\sigma) = L \neq ?$, then receive evidence σ inconsistent with L, and "hang on" to a refuted conjecture L until it changes its mind to L' at a future stage. This may lead to one extra mind change (from L to L') compared to the optimal number of mind changes that a learner may have achieved starting with evidence σ, for example by outputting ? until σ was observed.

A weaker requirement is that a learner Ψ has to be optimal for a subproblem $\mathcal{L}$ given σ only if $\Psi(\sigma)$ is consistent with σ. This leads us to the following definition. A conjecture $\Psi(\sigma)$ is **valid** for a sequence $\sigma \in \mathrm{SEQ}$ if $\Psi(\sigma) \neq ?$ and $\Psi(\sigma)$ is consistent with σ.

Definition 3. *A learner Ψ is* ***uniformly mind change optimal*** *for $\mathcal{L}$ given $\sigma \in \mathrm{SEQ}$ if there is a mind change counter c for Ψ such that (1) $c(\sigma) = \mathrm{MC}_{\mathcal{L}}(\sigma)$, and (2) for all data sequences $\tau \supseteq \sigma$, if $\Psi(\tau)$ is valid, then $c(\tau) = \mathrm{MC}_{\mathcal{L}}(\tau)$.*

We use the abbreviation "UMC-optimal" for "uniformly mind change optimal" (the terminology and intuition is similar to Kelly's in [15, 16]). A learner Ψ is simply UMC-optimal for $\mathcal{L}$ if Ψ is UMC-optimal given Λ.

Examples.
(1) In the COINIT problem, $\mathrm{MC}_{\mathcal{L}}(\Lambda) = \omega$, and $\mathrm{MC}_{\mathcal{L}}(\sigma) = \min(\mathrm{content}(\sigma))$ when $\mathrm{content}(\sigma) \neq \emptyset$. Since c_0 is a mind change counter for Ψ_{CO}, it follows that Ψ_{CO} is UMC-optimal. Any learner Ψ such that (1) $\Psi(\sigma) = \Psi_{\mathrm{CO}}(\sigma)$ if $\mathrm{content}(\sigma) \neq \emptyset$ and (2) $\Psi(\sigma) = \Psi(\sigma^-)$ if $\mathrm{content}(\sigma) = \emptyset$ is also UMC-optimal. (The initial conjecture $\Psi(\Lambda)$ is not constrained.)
(2) The learner Ψ_{LIN} is UMC-optimal. We will see that Ψ_{LIN} is the *only* learner that is both UMC-optimal and always outputs valid conjectures. Thus for the problem of inferring conservation laws, UMC-optimality coincides with the inferences of the physics community.

3 A Topological Characterization of Mind-Change Bounded Identifiability

Information-theoretical aspects of inductive inference have been studied by many learning theorists (e.g., [10] and [20]). As Jain et. al. observe [10, p.34]:

> Many results in the theory of inductive inference do not depend upon computability assumptions; rather, they are information theoretic in character. Consideration of noncomputable scientists thereby facilitates the analysis of proofs, making it clearer which assumptions carry the burden.

As an example, Angluin showed that her Condition 1 characterizes the indexed families of nonempty recursive languages inferable from positive data by computable learners [3, p.121] and that the noneffective version, Condition 2, is a necessary condition for inferability by computable learners.[1] Variants of Angluin's Condition 2 turn out to be both sufficient and necessary for various models of language identifiability by noncomputable learners ([20, Ch.2.2.2][10, Thm.3.26]). Information theoretic requirements such as Condition 2 constitute necessary conditions for computable learners, and are typically the easiest way to prove the unsolvability of some learning problems when they do apply. For example, Apsitis used the Baire topology on total recursive functions to show that $\mathbf{EX}_\alpha \neq \mathbf{EX}_{\alpha+1}$ [4, Sect.3]. On the positive side, if a sufficient condition for noneffective learnability is met, it often yields insights that lead to the design of a successful learning algorithm.

It has often been observed that point-set topology, one of the most fundamental and well-studied mathematical subjects, provides useful concepts for describing the information theoretic structure of learning problems [25, Ch.10], [21, 4, 14]. In particular, Apsitis investigated the mind change complexity of function learning problems in terms of the Baire topology [4]. He showed that Cantor's 1883 notion of accumulation order in a topological space [6] defines a natural ordinal-valued measure of complexity for function learning problems, and that accumulation order provides a lower bound on the mind change complexity of a function learning problem. We generalize Apsitis' use of topology to apply it to language collections. The following section briefly reviews the relevant topological concepts.

[1] Condition 2 characterizes BC-learnability for computable learners [5].

3.1 Basic Definitions in Point-set Topology

A **topological space** over a set X is a pair $(X, \mathcal{O})$, where $\mathcal{O}$ is a collection of subsets of X, called **open sets**, such that $\emptyset$ and X are in $\mathcal{O}$ and $\mathcal{O}$ is closed under arbitrary union and finite intersection. One way to define a topology for a set is to find a base for it. A **base** $\mathcal{B}$ for X is a class of subsets of X such that

1. $\bigcup \mathcal{B} = X$, and
2. for every $x \in X$ and any $B_1, B_2 \in \mathcal{B}$ that contain x, there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subseteq B_1 \cap B_2$.

For any base $\mathcal{B}$, the set $\{\bigcup \mathcal{C} : \mathcal{C} \subseteq \mathcal{B}\}$ is a topology for X [18, p.52]. That is, an open set is a union of sets in the base. Let $\mathcal{L}$ be a class of languages and $\sigma \in \mathrm{SEQ}$. We use $\mathcal{L}|\sigma$ to denote all languages in $\mathcal{L}$ that are consistent with σ (i.e., $\{L \in \mathcal{L} : L \text{ is consistent with } \sigma\}$); similarly $\mathcal{L}|D$ denotes the languages in $\mathcal{L}$ that include a given finite subset D. The next proposition shows that $\mathcal{B}_{\mathcal{L}} = \{\mathcal{L}|\sigma : \sigma \in \mathrm{SEQ}\}$ constitutes a base for $\mathcal{L}$.

Proposition 1. $\mathcal{B}_{\mathcal{L}} = \{\mathcal{L}|\sigma : \sigma \in \mathrm{SEQ}\}$ *is a base for* $\mathcal{L}$*; hence* $\mathcal{T}_{\mathcal{L}} = \{\bigcup \mathcal{S} : \mathcal{S} \subseteq \mathcal{B}_{\mathcal{L}}\}$ *is a topology for* $\mathcal{L}$*.*

The topology $\mathcal{T}_{\mathcal{L}}$ generalizes the **positive information topology** from recursion theory [24, p.186] if we consider the graphs of functions as languages (as in [10, Ch.3.9.2][20, Ch.2.6.2]).

Examples. For the language collection COINIT we have that $\mathrm{COINIT}\,|\{2,3\} = \{L_0, L_1, L_2\}$ and $\mathrm{COINIT}\,|\{0\} = \{L_0\}$.

In a topological space $(X, \mathcal{T})$, a point x is **isolated** if there is an open set $O \in \mathcal{T}$ such that $O = \{x\}$. If x is not isolated, then x is an **accumulation point of** X. Following Cantor [6], we define the **derived sets** using the concept of accumulation points.

Definition 4 (Cantor). *Let* $(X, \mathcal{T})$ *be topological space.*

1. *The* 0***-th derived set*** *of* X*, denoted by* $X^{(0)}$*, is just* X*.*
2. *For every successor ordinal* α*, the* α***-th derived set*** *of* X*, denoted by* $X^{(\alpha)}$*, is the set of all accumulation points of* $X^{(\alpha-1)}$*.*
3. *For every limit ordinal* α*, the set* $X^{(\alpha)}$ *is the intersection of all* β*-th derived sets, where* $\beta < \alpha$*. That is,* $X^{(\alpha)} = \bigcap_{\beta<\alpha} X^{(\beta)}$*.*

We give an example from the topology of the real plane that illustrates the geometrical intuitions behind the topological concepts.

Example. Let

$$A = \{(\frac{1}{n}, \frac{1}{m}) : n, m \in \mathbb{N}\} \cup \{(\frac{1}{n}, 0) : n \in \mathbb{N}\} \cup \{(0, \frac{1}{m}) : m \in \mathbb{N}\}$$

be a set of points in the real plane $\mathbb{R}^2$ with the standard topology. We use $\mathrm{iso}(X)$ to denote all isolated points in X. Then $\mathrm{iso}(A) = \{(\frac{1}{n}, \frac{1}{m}) : n, m \in \mathbb{N}\}$. Therefore

$$A^{(1)} = \{(\frac{1}{n}, 0) : n \in \mathbb{N}\} \cup \{(0, \frac{1}{m}) : m \in \mathbb{N}\}.$$

Similarly, we have $A^{(2)} = (0,0)$, and $A^{(3)} = \emptyset$.

In the topology $\mathcal{T}_{\mathcal{L}}$, a language L is an isolated point of $\mathcal{L}$ iff there is a finite subset $D \subseteq L$ such that the observation of D entails L (i.e., $\mathcal{L}|D = \{L\}$). The derived sets of $\mathcal{L}$ can be defined inductively as shown in Def. 4. Note if $\alpha < \beta$ then $\mathcal{L}^{(\alpha)} \supseteq \mathcal{L}^{(\beta)}$. It can be shown in set theory that there is an ordinal α such that $\mathcal{L}^{(\beta)} = \mathcal{L}^{(\alpha)}$, for all $\beta > \alpha$ [13]. In other words, there must be a fix point for the derivation operation. If $\mathcal{L}$ has an empty fix point, then we say $\mathcal{L}$ is **scattered** [18, p.78]. In a non-scattered space, the nonempty fixed point is called a **perfect kernel**.

The **accumulation order of a language** L in $\mathcal{L}$, denoted by $\mathrm{acc}_{\mathcal{L}}(L)$ is the maximum ordinal α such that $L \in \mathcal{L}^{(\alpha)}$; when $\mathcal{L}$ is fixed by context, we simply write $\mathrm{acc}(L) = \alpha$. The **accumulation order of a class of languages** $\mathcal{L}$, denoted by $\mathrm{acc}(\mathcal{L})$, is the supremum of the accumulation order of all languages in it. Therefore a language collection has an accumulation order if and only if it is scattered.[2]

Examples.

(1) The only isolated point in COINIT is $L_0 = \mathbb{N}$, for $\mathrm{COINIT}\,|\{0\} = \{L_0\}$. Therefore $\mathrm{COINIT}^{(1)} = \{L_i : i \geq 1\}$. Similarly L_1 is the only isolated point in $\mathrm{COINIT}^{(1)}$; hence $\mathrm{COINIT}^{(2)} = \{L_i : i \geq 2\}$. It is easy to verify that $\mathrm{COINIT}^{(n)} = \{L_i : i \geq n\}$. Therefore the accumulation order of language L_i in COINIT is i and the accumulation order of COINIT is $\omega = \sup \mathbb{N}$.

(2) In $\mathrm{LINEAR}_n = \{\text{linear subspaces of } \mathbb{Q}^n\}$, the only isolated point is $\mathbb{Q}^n$ itself: Let S be a set of n linearly independent points in $\mathbb{Q}^n$; then $\mathrm{LINEAR}_n\,|S = \{\mathbb{Q}^n\}$. Similarly every $(n-i)$-dimensional linear subspace of $\mathbb{Q}^n$ is an isolated point in $\mathrm{LINEAR}_n^{(i)}$. Therefore the accumulation order of LINEAR_n is n.

(3) In FIN, there is *no* isolated point. This is because for every finite subset S of $\mathbb{N}$, there are infinitely many languages in FIN that are consistent with S. Therefore FIN is a perfect kernel of itself and FIN has no accumulation order.

3.2 Accumulation Order Characterizes Mind Change Complexity

In this section we show that the accumulation order of a language collection $\mathcal{L}$ is an exact measure of its mind change complexity for (not necessarily effective) learners: if $\mathrm{acc}(\mathcal{L})$ is unbounded, then $\mathcal{L}$ is not identifiable with any ordinal mind change bound; and if $\mathrm{acc}(\mathcal{L}) = \alpha$, then $\mathcal{L}$ is identifiable with a mind change bound.[3]

In a language topology, accumulation order has two fundamental properties that we apply often. Let $\mathrm{acc}_{\mathcal{L}}(\sigma) \equiv \sup\{\mathrm{acc}_{\mathcal{L}}(L) : L \in \mathcal{L}|\sigma\}$; as usual, we omit the subscript in context. A language L in $\mathcal{L}$ **has the highest accumulation order given** σ if $\mathrm{acc}_{\mathcal{L}}(L) = \mathrm{acc}_{\mathcal{L}}(\sigma)$ and for every $L' \in \mathcal{L}|\sigma$, $L' \neq L$ implies $\mathrm{acc}_{\mathcal{L}}(L') < \mathrm{acc}_{\mathcal{L}}(L)$.

Lemma 1. *Let $\mathcal{L}$ be a scattered class of languages with bounded accumulation order.*

1. *For every language $L \in \mathcal{L}$, for every text T for L, there exists a time n such that L has the highest accumulation order given $T[n]$.*

---

[2] Accumulation order is also called scattering height, derived length, Cantor-Bendixson rank, or Cantor-Bendixson length [13].

[3] Necessary and sufficient conditions for finite mind change identifiability by learning *algorithms* appear in [19, 23].

2. *For any two languages $L_1, L_2 \in \mathcal{L}$ such that $L_1 \subset L_2$ it holds that* $\mathrm{acc}_{\mathcal{L}}(L_1) > \mathrm{acc}_{\mathcal{L}}(L_2)$.

Proof. Part 2 is immediate. Part 1: For contradiction, assume there is a text T for L such that for all n, $\mathcal{L}|(T[n])$ contains some language L' such that $\mathrm{acc}(L') \geq \mathrm{acc}(L) = \alpha$. Then L is an accumulation point of $\mathcal{L}^{(\alpha)}$, the subclass of $\mathcal{L}$ that contains all languages with accumulation order less than or equal to α. Therefore $\mathrm{acc}(L) \geq \alpha + 1$, which is a contradiction. □

We now establish the correspondence between mind change complexity and accumulation order: $\mathrm{MC}_{\mathcal{L}}(\sigma) = \mathrm{acc}_{\mathcal{L}}(\sigma)$.

Theorem 1. *Let $\mathcal{L}$ be a language collection and let σ be a finite data sequence. Then there is a learner Ψ that identifies $\mathcal{L}$ given σ with mind change bound $\alpha \iff \mathrm{acc}_{\mathcal{L}}(\sigma) \leq \alpha$.*

Proof. ($\Leftarrow$) We first prove by transfinite induction the auxiliary claim (*): if there is $L_\tau \in \mathcal{L}$ that has the highest accumulation order given data sequence τ, then there is a learner Ψ_τ and a counter c_τ such that (1) $\Psi_\tau(\tau) = L_\tau$, (2) Ψ_τ identifies $\mathcal{L}$ given τ, (3) c_τ is a mind change counter for Ψ_τ given τ, and (4) $c_\tau(\tau) = \mathrm{acc}(\mathcal{L}|\tau)$. Assume (*) for all $\beta < \alpha$ and consider $\alpha = \mathrm{acc}(\mathcal{L}|\tau)$. Note that (a) if $\tau^* \supset \tau$ and there is another language $L_{\tau^*} \neq L_\tau$ that has the highest accumulation order for τ^*, then $\mathrm{acc}(\mathcal{L}|\tau^*) < \mathrm{acc}(\mathcal{L}|\tau)$. Hence by inductive hypothesis, we may choose a learner Ψ_{τ^*} and c_{τ^*} with the properties (1)–(4). Now define Ψ_τ and c_τ as follows for $\tau' \supseteq \tau$.

1. $\Psi_\tau(\tau) := L_\tau$, and $c_\tau(\tau) := \alpha$.
2. if there is a τ^* such that: $\tau \subset \tau^* \subseteq \tau'$ and there is $L_{\tau^*} \neq L_\tau$ with the highest accumulation order for τ^*, then let τ^* be the least such sequence and set $\Psi_\tau(\tau') := \Psi_{\tau^*}(\tau')$, and $c_\tau(\tau') := c_{\tau^*}(\tau')$. (Intuitively, Ψ_τ follows Ψ_{τ^*} after τ^*).
3. otherwise $\Psi_\tau(\tau') := L_\tau$ and $c_\tau(\tau') := \alpha$.

(1) and (4) are immediate. We verify (2) and (3): Let $T \supset \sigma$ be a text for a target language $L \in \mathcal{L}$. If $L = L_\tau$, then Clause 2 never applies and Ψ_τ converges to L_τ on T without any mind changes after σ. Otherwise by Lemma. 1, there is a first stage n such that Clause 2 applies at $T[n]$. Then Ψ_τ converges to L by choice of $\Psi_{T[n]}$. Also, no mind change occurs at $T[n']$ for $|\sigma| < n' < n$. By (a) and definition of $c_\tau, c_{T[n]}$, we have that $c_\tau(T[n-1]) > c_{T[n]}(T[n])$. And c_τ follows $c_{T[n]}$ after stage n. This establishes (*).

Now we construct a learner Ψ as follows for all $\tau \supseteq \sigma$.

1. if there is a τ^* such that: $\sigma \subseteq \tau^* \subseteq \tau$ and there is L_{τ^*} with the highest accumulation order for τ^*, then let τ^* be the least such sequence and set $\Psi(\tau) := \Psi_{\tau^*}(\tau)$, and $c(\tau) := c_{\tau^*}(\tau)$. (Intuitively, Ψ follows Ψ_{τ^*} after τ^*).
2. Otherwise $\Psi(\tau) :=?$ and $c(\tau) := \mathrm{acc}(\mathcal{L}|\sigma)$.

We show that Ψ identifies $\mathcal{L}$ given σ. Let $L \in \mathcal{L}$ and let $T \supset \sigma$ be any text for L. Then by Lemma 1, there is a least time n such that some language L' has the highest accumulation order for $T[n]$. So the learner Ψ converges to L by choice of $\Psi_{T[n]}$. No mind change occurs at or before $T[n]$, and $\mathrm{acc}(\mathcal{L}|\sigma) \geq \mathrm{acc}(\mathcal{L}|T[n])$; this shows that c is a mind change counter for Ψ given σ.

($\Rightarrow$) Let Ψ be a learner that identifies $\mathcal{L}$ given σ and c is a mind change counter such that $c(\sigma) = \alpha$. We prove by transfinite induction that if $\mathrm{acc}(\sigma) > \alpha$, then c is not a mind change counter for $\mathcal{L}$. Assume the claim holds for all $\beta < \alpha$ and consider α. Suppose $\mathrm{acc}(\sigma) > \alpha$; then there is $L \in \mathcal{L}|\sigma$ such that $\mathrm{acc}(L) = \alpha + 1$. Case 1: $\Psi(\sigma) = L$. Then since L is a limit point of $\mathcal{L}^{(\alpha)}$, there is L' in $\mathcal{L}^{(\alpha)}$ such that $L' \neq L$ and $\mathrm{acc}(L') = \alpha$. Let $T' \supset \sigma$ be a text for L'. Since Ψ identifies L', there is a time $n > |\sigma|$ such that $\Psi(T'[n]) = L'$. Since $\Psi(T'[n]) \neq \Psi(\sigma)$ and $\Psi(\sigma) \neq ?$, this is a mind change of Ψ, hence $c(T'[n]) < c(\sigma)$. That is, $c(T'[n]) = \beta < \alpha$. On the other hand, since $\mathrm{acc}(L') = \alpha$, we have $\mathrm{acc}(T'[n]) > \beta$. By inductive hypothesis, c is not a mind change counter for Ψ. Case 2: $\Psi(\sigma) \neq L$. Let $T \supset \sigma$ be a text for L. Since Ψ identifies L, there is a time $n > |\sigma|$ such that $\Psi(T[n]) = L$. Since $c(T[n]) \leq c(\sigma) = \alpha$ and $\mathrm{acc}(T[n]) > \alpha$, as in Case 1, c is not a mind change counter for Ψ. □

Corollary 1. *Let $\mathcal{L}$ be a class of languages. Then there exists a mind-change bound for $\mathcal{L}$ if and only if $\mathcal{L}$ is scattered in the topology $\mathcal{T}_{\mathcal{L}}$.*

4 Necessary and Sufficient Conditions for Uniformly Mind Change Optimal learners

The goal of this section is to characterize the behaviour of uniformly mind-change optimal learners. These results allow us to design mind change optimal learners and to prove their optimality. The next definition specifies the key property of uniformly MC-optimal learners.

Definition 5. *A learner Ψ is **order-driven** given σ if for all finite data sequences $\tau, \tau' \in \mathrm{SEQ}(\mathcal{L})$ such that $\sigma \subseteq \tau \subset \tau'$: if (1) $\tau = \sigma$ or $\Psi(\tau)$ is valid for τ, and (2) $\mathrm{acc}_{\mathcal{L}}(\tau) = \mathrm{acc}_{\mathcal{L}}(\tau')$, then Ψ does not change its mind at τ'.*

Informally, a learner Ψ is order-driven if once Ψ makes a valid conjecture $\Psi(\tau)$ at τ, then Ψ "hangs on" to $\Psi(\tau)$ at least until the accumulation order drops at some sequence $\tau' \supset \tau$, that is, $\mathrm{acc}(\tau') < \mathrm{acc}(\tau)$. Both the learners Ψ_{CO} and Ψ_{LIN} are order-driven given Λ.

A data sequence σ is **topped** if there is a language $L \in \mathcal{L}$ consistent with σ such that $\mathrm{acc}_{\mathcal{L}}(L) = \mathrm{acc}_{\mathcal{L}}(\sigma)$. Note that if $\mathrm{acc}_{\mathcal{L}}(\sigma)$ is a successor ordinal (e.g., finite), then σ is topped. All data sequences in $\mathrm{SEQ}(\mathrm{LINEAR}_n)$ are topped. In COINIT, the initial sequence Λ is *not* topped. As the next proposition shows, if σ is topped, the conjecture $\Psi(\sigma)$ of a UMC-optimal learner Ψ is *highly constrained*: either $\Psi(\sigma)$ is not valid, or else $\Psi(\sigma)$ must uniquely have the highest accumulation order in $\mathcal{L}|\sigma$.

Proposition 2. *Let $\mathcal{L}$ be a language collection such that $\mathrm{acc}_{\mathcal{L}}(\sigma) = \alpha$ for some ordinal α and data sequence σ. Suppose that learner Ψ is uniformly mind change optimal and identifies $\mathcal{L}$ given σ. Then*

1. *Ψ is order-driven given σ.*

2. *for all data sequences $\tau \supseteq \sigma$, if τ is topped and $\Psi(\tau)$ is valid for τ, then $\Psi(\tau)$ is the unique language with the highest accumulation order for τ.*

Proof Outline. Clause 1. Suppose that $\tau = \sigma$ or that $\Psi(\tau)$ is valid for $\tau \supset \sigma$; then $c(\tau) = \mathrm{acc}(\tau)$. If Ψ changes its mind at τ' when $\mathrm{acc}(\tau') = \mathrm{acc}(\tau)$, then $c(\tau') < c(\tau) = \mathrm{acc}(\tau')$. Hence by Theorem 1, c is not a mind change counter for Ψ.

Clause 2. Suppose for reductio that $\Psi(\tau)$ is valid for τ but $\Psi(\tau)$ does not have the highest accumulation order for τ. Then there is a language $L \in \mathcal{L}|\tau$ such that (1) $\mathrm{acc}(L) = \mathrm{acc}(\tau)$, and (2) $\mathrm{acc}(L) \geq \mathrm{acc}(\Psi(\tau))$, and (3) $L \neq \Psi(\tau)$. Choose any text $T \supset \tau$ for L. Since Ψ identifies $\mathcal{L}$, there is an $n > |\tau|$ such that $\Psi(T[n]) \neq \Psi(\tau)$ and $\mathrm{acc}(T[n]) = \mathrm{acc}(\tau)$. Hence Ψ is not order-driven. □

To illustrate, in COINIT, since the initial sequence Λ is not topped, Prop. 2 does not restrict the conjectures of UMC-optimal learners at Λ.

A learner Ψ is **regular** given σ if for all data sequences $\tau \supset \sigma$, if Ψ changes its mind at τ, then $\Psi(\tau)$ is valid. Intuitively, there is no reason for a learner Ψ to change its conjecture to an invalid one. The learners Ψ_{COINIT} and Ψ_{LIN} are regular. According to Prop. 2, being order-driven is necessary for a UMC-optimal learner. The next proposition shows that for regular learners, this property is sufficient as well.

Proposition 3. *Let $\mathcal{L}$ be a language collection such that $\mathrm{acc}(\sigma) = \alpha$ for some ordinal α and data sequence σ. If a learner Ψ identifies $\mathcal{L}$ and is regular and order-driven given σ, then Ψ is uniformly mind change optimal given σ.*

Proof. Let Ψ be regular and order-driven given σ. Define a counter c as follows for σ and $\tau \supset \sigma$.

(1) $c(\sigma) = \mathrm{acc}(\sigma)$.

(2) $c(\tau) = \mathrm{acc}(\tau)$ if $\Psi(\tau)$ is valid for τ.

(3) $c(\tau) = c(\tau^-)$ if $\Psi(\tau)$ is not valid for τ.

Clearly $c(\tau) = \mathrm{acc}(\tau)$ if $\Psi(\tau)$ is valid for τ. So it suffices to show that c is a mind change counter for Ψ. Let Ψ change its mind at $\tau \supset \sigma$. Then since Ψ is regular given σ, we have that $\Psi(\tau)$ is valid for τ and hence (a) $c(\tau) = \mathrm{acc}(\tau)$.

Case 1: There is a time n such that (1) $|\sigma| \leq n \leq \mathrm{lh}(\tau^-)$, where $\mathrm{lh}(\tau^-)$ is the length of τ^-, and (2) $\Psi(\tau^-[n])$ is valid for $\tau^-[n]$. WLOG, let n be the greatest such time. Then by the definition of c, we have that (b) $c(\tau^-[n]) = c(\tau^-)$. Since $\tau^-[n] \subset \tau$, and Ψ changes its mind at τ, and Ψ is order-driven, it follows that (c) $\mathrm{acc}(\tau^-[n]) > \mathrm{acc}(\tau)$. Also, by (2), we have that (d) $c(\tau^-[n]) = \mathrm{acc}(\tau^-[n])$. Combining (a), (b), (c) and (d), it follows that $c(\tau^-) > c(\tau)$.

Case 2: There is no time n such that $|\sigma| \leq n \leq \mathrm{lh}(\tau^-)$and (2) $\Psi(\tau^-[n])$ is valid for $\tau^-[n]$. Then by definition of c, we have that (e) $c(\tau^-) = \mathrm{acc}(\sigma)$. And since Ψ is order-driven given σ, (f) $\mathrm{acc}(\sigma) > \mathrm{acc}(\tau)$. Combining (a), (e), and (f), we have that $c(\tau^-) > c(\tau)$.

So in either case, if Ψ changes its mind at $\tau \supset \sigma$, then $c(\tau^-) > c(\tau)$, which establishes that c is a mind change counter for Ψ given σ. Hence Ψ is UMC-optimal given σ. □

In short, Propositions 2 and 3 show that being order-driven is the key property of a uniformly mind change optimal learner.

Examples.
(1) In COINIT, for any data sequence $\sigma \in \mathrm{SEQ}$ such that $\mathrm{content}(\sigma) \neq \emptyset$, we have that $\mathcal{L}|\sigma$ is topped and there is a unique language $L(\sigma)$ with the highest accumulation order. Since $\Psi_{\mathrm{CO}}(\sigma) = L(\sigma)$ whenever $\mathrm{content}(\sigma) \neq \emptyset$, the learner $\Psi_{\mathrm{CO}}(\sigma)$ is order-driven and regular, and hence a UMC-optimal learner for COINIT by Prop. 3. But Ψ_{CO} is not the unique UMC-optimal learner: Define a modified learner Ψ_0^k by setting $\Psi_0^k(\sigma) := L_k$ if $\mathrm{content}(\sigma) = \emptyset$, and $\Psi_0^k(\sigma) := \Psi_{\mathrm{CO}}(\sigma)$ otherwise. Any such learner Ψ_0^k is a valid uniformly MC-optimal learner.
(2) Since LINEAR_n is finite, it is a topped language collection. In fact, for all data sequences σ, the language with the highest accumulation order is given by $\mathrm{span}(\mathrm{vectors}(\sigma))$. Thus the learner Ψ_{LIN} is the unique uniformly MC-optimal learner for LINEAR_n such that $\Psi_{\mathrm{LIN}}(\sigma)$ is valid for all data sequences $\sigma \in \mathrm{SEQ}(\mathrm{LINEAR}_n)$.

5 Effective Uniformly Mind Change Optimal Learning

It is straightforward to computationally implement the learners Ψ_{CO} and Ψ_{LIN}. These learners have the feature that whenever they produce a conjecture L on data σ, the language L is the $\subseteq$-minimum among all languages consistent with σ. It follows immediately from Clause 2 of Lemma. 1 that Ψ_{CO} and Ψ_{LIN} always output an order-maximizing hypothesis (the language uniquely having the highest accumulation order). For many problems, e.g., COINIT and LINEAR_n, a language has the highest accumulation order iff it is the $\subseteq$-minimum. For such a language collection $\mathcal{L}$, if we can compute the $\subseteq$-minimum, a UMC-optimal learning algorithm for $\mathcal{L}$ can be constructed on the model of Ψ_{CO} and Ψ_{LIN}. However, these conditions are much stronger than necessary in general. In general, it suffices that we can *eventually* compute a $\subseteq$-minimum along any text. We illustrate this point by specifying a UMC-optimal learning algorithm for P_1, the languages defined by Angluin's well-known one-variable patterns [2, p.48].

Let X be a set of variable symbols and let Σ be a finite alphabet of at least two constant symbols (e.g., $0, 1, \ldots, n$). A **pattern**, denoted by p, q etc., is a finite non-null sequence over $X \cup \Sigma$. If a pattern contains exactly one distinct variable, then it is a **one-variable pattern** (e.g., `x01` or `0x00x1`). Following [2], we denote the set of all one-variable patterns by P_1. A **substitution** θ replaces x in a pattern p by another pattern. For example, $\theta = [\mathtt{x}/\mathtt{0}]$ maps the pattern `xx` to the pattern `00` and $\theta' = [\mathtt{x}/\mathtt{xx}]$ maps the pattern `xx` to the pattern `xxxx`. Substitutions give rise to a partial order $\preceq$ over all patterns. Let p and q be two patterns. We define $p \preceq q$ if there is a substitution θ such that $p = q\theta$. The **language generated by a pattern** p, denoted by $L(p)$, is the set $\{q \in \Sigma^* : q \preceq p\}$.

Angluin described an algorithm that, given a finite set S of strings as input, finds the set of one-variable patterns descriptive of S, and then (arbitrarily) selects one with the maximum length [2, Th.6.5]. A one-variable pattern p is **descriptive of a sample** S if $S \subseteq L(p)$ and for every one-variable pattern q such that $S \subseteq L(q)$, the language $L(q)$ is not a proper subset of $L(p)$ [2, p.48]. To illustrate, the pattern `1x` is descriptive of the samples $\{\mathtt{10}\}$ and $\{\mathtt{10}, \mathtt{11}\}$, the pattern `x0` is descriptive of the samples $\{\mathtt{10}\}$ and $\{\mathtt{10}, \mathtt{00}\}$, and the pattern `x` is descriptive of the sample $\{\mathtt{10}, \mathtt{00}, \mathtt{11}\}$. We give an example (summarised in Fig. 1) to show that Angluin's algorithm is not a mind-change optimal learner. Let $\mathtt{x}$ be the target pattern and consider the text $T = \langle \mathtt{10}, \mathtt{00}, \mathtt{11}, \mathtt{0}, \ldots \rangle$ for $L(\mathtt{x})$. Let us write $P_1|S$ for the set of one-variable patterns consistent with a sample S. Then $P_1|\{\mathtt{10}\} = \{\mathtt{1x}, \mathtt{x0}, \mathtt{x}\}$, $P_1|\{\mathtt{10}, \mathtt{00}\} = \{\mathtt{x0}, \mathtt{x}\}$, $P_1|\{\mathtt{10}, \mathtt{11}\} = \{\mathtt{1x}, \mathtt{x}\}$ and $P_1|\{\mathtt{10}, \mathtt{00}, \mathtt{11}\} = \{\mathtt{x}\}$. The accumulation orders of these languages are determined as follows:

1. $\mathrm{acc}_{P_1}(L(\mathtt{x})) = 0$ since $L(\mathtt{x})$ is isolated; so $\mathrm{acc}_{P_1}(\langle \mathtt{10}, \mathtt{00}, \mathtt{11} \rangle) = 0$.
2. $\mathrm{acc}_{P_1}(L(\mathtt{1x})) = 1$ since $P_1|\{\mathtt{10}, \mathtt{11}\} = \{\mathtt{1x}, \mathtt{x}\}$; so $\mathrm{acc}_{P_1}(\langle \mathtt{10}, \mathtt{11} \rangle) = 1$.
3. $\mathrm{acc}_{P_1}(L(\mathtt{x0})) = 1$ since $P_1|\{\mathtt{10}, \mathtt{00}\} = \{\mathtt{x0}, \mathtt{x}\}$; so $\mathrm{acc}_{P_1}(\langle \mathtt{10}, \mathtt{00} \rangle) = 1$.

Also, we have $\mathrm{acc}_{P_1}(\langle \mathtt{10} \rangle) = 1$. Since for $T[1] = \langle \mathtt{10} \rangle$, the one-variable patterns $\mathtt{1x}$ and $\mathtt{x0}$ are both descriptive of $\{\mathtt{10}\}$, Angluin's learner M_A conjectures either $\mathtt{1x}$ or $\mathtt{x0}$; suppose $M_A(\langle \mathtt{10} \rangle) = \mathtt{1x}$. Now let c_A be any mind change counter for M_A. Since $\mathtt{1x}$ is consistent with $\langle \mathtt{10} \rangle$, UMC-optimality requires that $c_A(\langle \mathtt{10} \rangle) = \mathrm{acc}_{P_1}(\langle \mathtt{10} \rangle)= 1$. The next string $\mathtt{00}$ in T refutes $\mathtt{1x}$, so M_A changes its mind to $\mathtt{x0}$ (i.e., $M_A(T[2]) = \mathtt{x0}$), and $c_A(\langle \mathtt{10}, \mathtt{00} \rangle) = 0$. However, M_A changes its mind again to pattern $\mathtt{x}$ on $T[3] = \langle \mathtt{10}, \mathtt{00}, \mathtt{11} \rangle$, so c_A is not a mind change counter for M_A, and M_A is not UMC-optimal. In short, after the string $\mathtt{10}$ is observed, it is possible to identify the target one-variable pattern with one more mind change, but M_A requires two.

The issue with M_A is that M_A changes its mind on sequence $\langle \mathtt{10}, \mathtt{00} \rangle$ even though $\mathrm{acc}_{P_1}(\langle \mathtt{10} \rangle) = \mathrm{acc}_{P_1}(\langle \mathtt{10}, \mathtt{00} \rangle) = 1$, so M_A is not order-driven and hence Proposition 2 implies that M_A is not UMC-optimal. Intuitively, an order-driven learner has to wait until the data decide between the two patterns $\mathtt{1x}$ and $\mathtt{x0}$. As Proposition 3 indicates, we can design a UMC-optimal learner M for P_1 by "procrastinating" until there is a pattern with the highest accumulation order. For example on text T our UMC-optimal learner M makes the following conjectures: $M(\langle \mathtt{10} \rangle) =?$, $M(\langle \mathtt{10}, \mathtt{00} \rangle) = \mathtt{x0}$, $M(\langle \mathtt{10}, \mathtt{00}, \mathtt{11} \rangle) = \mathtt{x}$.

Text T	:	10	00	11	0	$\cdots$
Stage n	:	1	2	3	4	$\cdots$
Patterns consistent with $T[n]$	:	1x, x0, x	x0, x	x	x	$\cdots$
Patterns descriptive of $T[n]$	:	1x, x0	x0	x	x	$\cdots$
Accumulation order of $T[n]$	:	1	1	0	0	$\cdots$
Output of Angluin's learner M_A	:	1x	x0	x	x	$\cdots$
Output of a UMC-optimal learner M	:	?	x0	x	x	$\cdots$

Fig. 1. An illustration of why Angluin's learning algorithm for one-variable patterns is not uniformly mind change optimal.

The general specification of the UMC-optimal learning algorithm M is as follows. For a terminal $a \in \Sigma$ let $p^a \equiv p[x/a]$. The proof of [2, Lemma 3.9] shows that if q

is a one-variable pattern such that $L(q) \supseteq \{p^a, p^b\}$ for two distinct terminals a, b, then $L(q) \supseteq L(p)$. Thus a UMC-optimal learning algorithm M can proceed as follows.

1. Set $M(\Lambda) :=?$.
2. Given a sequence σ with $S := \text{content}(\sigma)$, check (*) if there is a one-variable pattern p consistent with σ such that $S \supseteq \{p^a, p^b\}$ for two distinct terminals a, b. If yes, output $M(\sigma) := p$. If not, set $M(\sigma) := M(\sigma^-)$.

Since there are at most finitely many patterns consistent with σ, the check (*) is effective. In fact, (*) and hence M can be implemented so that computing $M(\sigma)$ takes time linear in $|\sigma|$. Outline: Let $m = \min\{|s| : s \in S\}$. Let S^m be the set of strings in S of length m. Define $p_S(i) := a$ if $s(i) = a$ for all $s \in S^m$, and $p_S(i) := \mathtt{x}$ otherwise for $1 \leq i \leq m$. For example, $p_{\{\mathtt{10,11,111}\}} = \mathtt{1x}$ and $p_{\{\mathtt{10,01}\}} = \mathtt{x}$. Then check for all $s \in S$ if $s \in L(p_S)$. For a one-variable pattern, this can be done in linear time because $|\theta(\mathtt{x})|$, the length of $\theta(\mathtt{x})$, must be $\frac{|s| - term(p_S)}{|p_S| - term(p_S)}$ where $term(p_S)$ is the number of terminals in p_S. For example, if $s = \mathtt{111}$ and $p_S = \mathtt{1x}$, then $|\theta(\mathtt{x})|$ must be 2. If p_S is consistent with S, then there are distinct $a, b \in \Sigma$ such that $\{p^a, p^b\} \subseteq S$. Otherwise no pattern p of length m is consistent with S and hence (*) fails.

6 Summary and Future Work

The topic of this paper was learning with bounded mind changes. We applied the classic topological concept of accumulation order to characterize the mind change complexity of a learning problem: A language collection $\mathcal{L}$ is identifiable by a learner (not necessarily computable) with α mind changes iff the accumulation order of $\mathcal{L}$ is at most α. We studied the properties of uniformly mind change optimal learners: roughly, a learner Ψ is uniformly mind change optimal if Ψ realizes the best possible mind change bound not only in the entire learning problem, but also in subproblems that arise after observing some data. The characteristic property of UMC-optimal learners is that they output languages with maximal accumulation order. Thus analyzing the accumulation order of a learning problem is a powerful guide to constructing mind change efficient learners. We illustrated these results in several learning problems such as identifying a linear subspace and a one-variable pattern. For learning linear subspaces, the natural method of conjecturing the least subspace containing the data is the only mind change optimal learner that does not "procrastinate" (i.e., never outputs ? or an inconsistent conjecture). Angluin's algorithm for learning a one-variable pattern is not UMC-optimal; we described a different UMC-optimal algorithm for this problem.

We outline several avenues for future work. The next challenge for pattern languages is to find a UMC-optimal algorithm for learning a general pattern with arbitrarily many variables. An important step towards that goal would be to determine the accumulation order of a pattern language $L(p)$ in the space of pattern languages. Another application is the design of UMC-optimal learners for logic programs. For example, Jain and Sharma have examined classes of logic programs that can be learned with bounded mind changes using explorer trees [12]. Do explorer trees lead to mind change optimal learning algorithms?

There are a number of open issues for the general theory of UMC-optimal learning. The proof of Theorem 1 shows that if there is any general learner that solves a learning problem $\mathcal{L}$ with α mind changes, then there is a UMC-optimal general learner for $\mathcal{L}$. However, this may well not be the case for effective learning algorithms: Is there a language collection $\mathcal{L}$ such that there is a computable learner M that identifies $\mathcal{L}$ with α mind changes, but there is no computable UMC-optimal learner for $\mathcal{L}$? Such a separation result would show that for computable learners, UMC-optimality defines a new class of learning problems.

As the example of one-variable patterns shows, there can be a trade-off between time efficiency and producing consistent conjectures, on the one hand, and the procrastination that minimizing mind changes may require on the other (see Sect. 5). We would like to characterize the learning problems for which this tension arises, and how great the trade-off can be.

Another project is to relate the topological concept of accumulation order to other well-known structural properties of a language collection $\mathcal{L}$. For example, it can be shown that if $\mathcal{L}$ has unbounded accumulation order (i.e., if $\mathcal{L}$ contains a nonempty perfect subset), then $\mathcal{L}$ has infinite elasticity, as defined in [29, 22]. Also, we can show that accumulation order corresponds to intrinsic complexity as defined in [7, 11], in the following sense: If $\mathcal{L}_1$ is weakly reducible to $\mathcal{L}_2$, then the accumulation order of $\mathcal{L}_2$ is at least as great as the accumulation order of $\mathcal{L}_1$. It follows immediately that COINIT $\not\leq_{weak}$ SINGLE, where SINGLE is the class of all singleton languages and has accumulation order 0, and FIN $\not\leq_{weak}$ COINIT, two results due to Jain and Sharma [11].

In sum, uniform mind change optimality guides the construction of learning algorithms by imposing strong and natural constraints; and the analytical tools we established for solving these constraints reveal significant aspects of the fine structure of learning problems.

Acknowledgments

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. We would like to thank the anonymous COLT referees for their comments and suggestions.

References

1. A. Ambainis, S. Jain, and A. Sharma. Ordinal mind change complexity of language identification. *Theor. Comput. Sci.*, 220(2):323–343, 1999.
2. D Angluin. Finding patterns common to a set of strings. *J. Comput. Syst. Sci.*, 21(1):46–62, 1980.
3. D Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135, 1980.
4. K. Apsitis. Derived sets and inductive inference. In S. Arikawa and K. P. Jantke, editors, *Proceedings of ALT 1994*, pages 26–39. Springer, Berlin, Heidelberg, 1994.
5. G. Baliga, J. Case, and S. Jain. The synthesis of language learners. *Information and Computation*, 152:16–43, 1999.

6. G. Cantor. Grundlagen einer allgemeinen Mannigfaltigkeitslehre. In William Ewald, editor, *From Kant to Hilbert*, volume 2, pages 878–920. Oxford Science Publications, 1996.
7. R. Freivalds, E. Kinber, and C. H. Smith. On the intrinsic complexity of learning. *Inf. Comput.*, 123(1):64–71, 1995.
8. R. Freivalds and C. H. Smith. On the role of procrastination in machine learning. *Inf. Comput.*, 107(2):237–271, 1993.
9. E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.
10. S. Jain, D. Osherson, J. S. Royer, and A. Sharma. *Systems That Learn*. M.I.T. Press, 2 edition, 1999.
11. S. Jain and A. Sharma. The intrinsic complexity of language identification. *J. Comput. Syst. Sci.*, 52(3):393–402, 1996.
12. S. Jain and A. Sharma. Mind change complexity of learning logic programs. *TCS*, 284(1):143–160, 2002.
13. A. J. Jayanthan. Derived length for arbitrary topological spaces. *International Journal of Mathematics and Mathematical Sciences*, 15(2):273–277, 1992.
14. K. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, 1996.
15. K. Kelly. Efficient convergence implies Ockham's Razor. In *Proceedings of the 2002 International Workshop on Computation Models of Scientific Reasoning and Applications*, pages 24–27, 2002.
16. K. Kelly. Justification as truth-finding efficiency: How ockham's razor works. *Minds and Machines*, 14(4):485–505, 2004.
17. S. Kocabas. Conflict resolution as discovery in particle physics. *Machine Learning*, 6:277–309, 1991.
18. K. Kuratowski. *Topology*, volume 1. Academic Press, 1966. Translated by J. Jaworowski.
19. S. Lange and T. Zeugmann. Language learning with a bounded number of mind changes. In *Symposium on Theoretical Aspects of Computer Science*, pages 682–691, 1993.
20. E. Martin and D. N. Osherson. *Elements of Scientific Inquiry*. The MIT Press, Cambridge, Massachusetts, 1998.
21. E. Martin, A. Sharma, and F. Stephan. Learning, logic, and topology in a common framework. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, pages 248–262. Springer-Verlag, 2002.
22. T. Motoki, T. Shinohara, and K. Wright. The correct definition of finite elasticity: corrigendum to identification of unions. In *Proceedings of COLT 1991*, page 375. Morgan Kaufmann Publishers Inc., 1991.
23. Y. Mukouchi. Inductive inference with bounded mind changes. In S. Doshita, K. Furukawa, K. P. Jantke, and T. Nishida, editors, *Proceedings of ALT 1992*, pages 125–134. Springer, Berlin, Heidelberg, 1993.
24. P. Odifreddi. *Classical Recursion Theory*. North-Holland, October 1999.
25. D. N. Osherson, M. Stob, and S. Weinstein. *Systems that learn: an introduction to learning theory for cognitive and computer scientists*. MIT Press, 1986.
26. O. Schulte. Automated discovery of conservation principles and new particles in particle physics. Manuscript submitted to *Machine Learning*, 2005.
27. R. Valdés-Pérez. Algebraic reasoning about reactions: Discovery of conserved properties in particle physics. *Machine Learning*, 17:47–67, 1994.
28. R. Valdés-Pérez. On the justification of multiple selection rules of conservation in particle physics phenomenology. *Computer Physics Communications*, 94:25–30, 1996.
29. K. Wright. Identification of unions of languages drawn from an identifiable class. In *Proceedings of the second annual workshop on Computational learning theory*, pages 328–333. Morgan Kaufmann Publishers Inc., 1989.