Getting Order Independence

in Incremental Learning

Antoine Cornuéjols

Equipe Inférence et Apprentissage
Laboratoire de Recherche en Informatique (LRI), UA 410 du CNRS
Université de Paris-sud, Orsay
Bâtiment 490, 91405 ORSAY (France)
email (UUCP) : antoine@lri.lri.fr

Abstract[*] . It is empirically known that most incremental learning systems are order dependent, i.e. provide results that depend on the particular order of the data presentation. This paper aims at uncovering the reasons behind this, and at specifying the conditions that would guarantee order independence. It is shown that both an optimality and a storage criteria are sufficient for ensuring order independence. Given that these correspond to very strong requirements however, it is interesting to study necessary, hopefully less stringent, conditions. The results obtained prove that these necessary conditions are equally difficult to meet in practice.

Besides its main outcome, this paper provides an interesting method to transform an history dependent bias into an history independent one.

1   Introduction

Ordering effects in Incremental Learning have been widely mentioned in the literature without, however, being the subject of much specific study except for some rare pioneering works [1,4,9, and, incidentally, in 2][1]. In short, ordering effects are observed when, given a collection of data (e.g. examples in inductive concept learning), different ordered sequences of these data lead to different learning results. In this respect, ordering of data

---

[*] This paper appeared in the Proc. of the *European Conference on Machine Learning* (ECML-93), Vienna, April, 5-8, 1993.

[1] See also the AAAI Technical Report corresponding to the recent AAAI Spring Symposium (March 23-25, 1993, Stanford University) devoted to "Training Issues in Incremental Learning".

therefore seems to be equivalent to a preference bias that makes a choice among all the models or hypotheses that the learning system could reach given the collection of data (that is the models that would be obtained had the collection of data been presented in every possible orders). Hence, ordering undoubtedly amounts to some additional knowledge supplied to the system. This is why teachers have some value: by selecting suitable pedagogical presentations of the material to be learned they provide further knowledge that hopefully helps the learning process.

Learning without some bias that allows the reduction of the search space for the target concept or model is impossible except in the crudest form of rote learning. When looking more closely, it is usual to distinguish between :

> - *representation bias* : where the search space is constrained because all partitions of the example space can not be expressed in the hypothesis space considered by the system (this is the basis for inductive generalization and is the main topic of current Machine Learning theory [12]), and

> - *preference bias* : which dictates which subspace should be preferred in the search space (e.g. prefer simple hypotheses over more complex ones) (this type of bias has been much less studied because it touches on procedural aspects instead of declarative ones only).

Because ordering of inputs allows one to favor some models over some others, it seems to amount to a preference bias that chooses between competing hypotheses. In spite of this resemblance however, there is a deep difference with the biases generally discussed in Machine Learning. Indeed, with ordering effects, one *observes* the preference but cannot pinpoint directly where

in the learning system it lies and how it works. This is in contrast with what is considered classically as a bias, where one can identify *operational* constraints _e.g. isolate representation constraints or procedures for choice between hypotheses. Thus we use the term <u>global preference bias</u> to denote preference among models due to ordering effects *after* a sequence of inputs has been observed, and the term <u>local preference bias</u> to denote the local choice strategy followed by the system when at each learning step it must choose to follow some paths and discard others.

Two questions then immediately come up :

1- *What is the relationship between a global preference bias and a local one ?*

2- *What is the relationship between a global preference bias that is observed or aimed at and a corresponding teaching strategy that specifies the order of inputs ? In other words, how to design a teaching strategy so as to get a certain global preference bias ?*

The second question is related to the recently introduced concept of teachability [5,8,15,16,17] and the not so recent concern for good training sequences [19,20]. However, it differs in a fundamental point in that, in the former the problem is essentially the determination of good examples to speed up learning, whereas in our setting we assume that the collection of instances is given a priori and our only degree of freedom lies in the choice of a good ordering. Additionally, researchers in the teachability concept have not been interested with the idea of guiding the learner toward some preferred model and away from others, they seek to characterize the learnability of concept classes irrelevant of particular preferences within the classes. Keeping in mind these differences, the emphasis on providing additional knowledge to the learner through an educational strategy is the same.

In order to answer these questions, and particularly the first one, it is necessary to determine the causes of the ordering effects.

2 Causes of ordering effects

It is instructive to look at incremental learners that are NOT order dependent, like the candidate elimination (CE) algorithm in Version Space [10,11], ID5R [18], or systems that are not usually considered as learning systems but could be, such as TMS [3] or some versions of the Bayesian Inference nets of Pearl [13]. They all have in common that they do not forget any information present in the input data. Thus, even when they make a choice between alternative hypotheses, like ID5 or TMS and unlike the CE algorithm, they keep enough information to be able to compare all potential competing models so as to select the best one at any moment, and change their mind if needed. They are therefore equivalent to non-incremental learning systems that get all the data at once and focus on the best hypothesis given the information supplied.

To sum up, order independent incremental learners *(i)* are able to focus on a optimal hypothesis when they have to choose among the current potential ones; and *(ii)* they do keep enough informations so as to not forget any potential hypothesis. If one or both of these properties is lacking, then incremental learning is prone to be order dependent. A closer look at each of these property in turn will help to see why.

(i) Optimality vs. non optimality.

Since the influential thesis and papers of Mitchell [10,11], is has become commonplace to consider learning as a search process in a concept or solution space. Whatever the form (generalization, explanation,...) and the constraints on the hypothesis space (e.g. representation language bias), the learner is searching the best solution in this space given the data at hand. In order to compare the solutions in the search space, it is possible to imagine that the learner evaluates and grades each one of them and then chooses the top one. Of course, the grade and therefore the rank of these solutions can be modified if the data are changed.

Because incremental learners usually function by adapting a current solution (or a small solution set) to cope with new informations, they proceed typically in a hill-climbing fashion (see figure 1) open to the draw-back of missing the global optimum solution in favor of local ones.
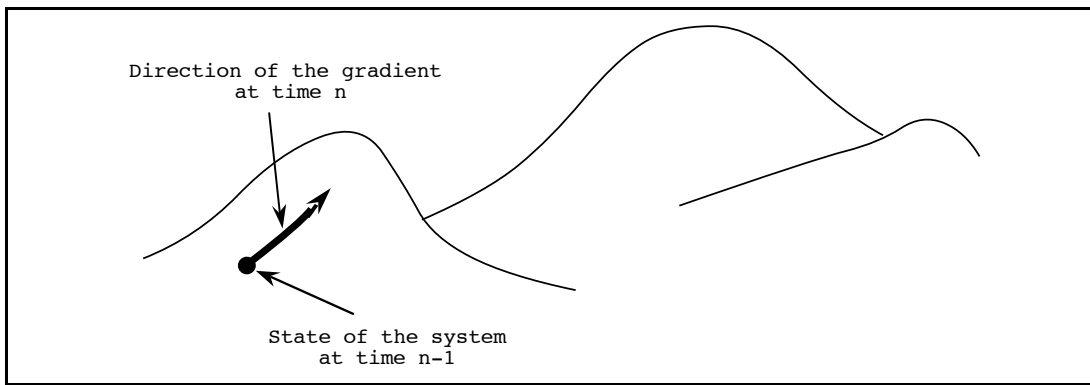
Fig. 1. At time n-1, the system is in a state corresponding to a given hypothesis. At time n, with a new arriving piece of data, the value of each hypothesis is re-evaluated and the system follows the direction of greatest gradient to reach a new state.

As underlined above, one must realize that in contrast to the common optimization problems where the optimums and the whole topology of the search space are set once and for all before the search starts, in incremental learning the topology is changed with each new input and the solutions in the hypothesis space are therefore open to re-evaluation, new optimums replacing old ones. Because of this, incremental learning wandering from one local and temporary optimum to the next can be order dependent unless it reaches the global current optimum at each step that is with each arriving data[2]. In that case of course, at the end of the training period and given that all the training data have been observed, the system would reach the global optimum for this set of data regardless of its ordering during learning.

In fact, it is easy to see that there is a second caveat in addition to the requirement of finding the global optimum at each time.

(ii) To forget or not to forget.

Finding the optimal solution at each step is operating only to the extent that the set of all possible solutions is brought to consideration; if only part of it is available then the whole optimization process is bound to be sub-optimal. Now, beside adapting the current solution to new constraints expressed under the form of new data,

incremental learners because they entertain only some preferred hypothesis or model, can also discard in the process part of the information present in the past inputs. If this forgetting is dependent upon the current state of the system, then it follows that it may also be dependent upon the ordering of the inputs. Therefore the resulting learning may be order dependent.

It is important to note that these two reasons if often tied (this is because one keeps only a current solution for adaptation that one is tempted to forget informations about other possibilities) are nonetheless independent in principle and can be observed and studied separately. For reasons that will be exposed shortly, we will concentrate on the forgetting aspect of incremental learning and will accordingly assume, by way of an adequate framework, that the best alternatives among the available ones can always be chosen by the learner.

Forgetting of information lies therefore at the heart of order dependence in incremental learning. But *forgetting can take two faces*. In the first one, information present in the input data is lost, meaning that the current hypothesis space considered by the learner is underconstrained. In the second one, by contrast, what is lost are potential alternatives to the current preferred hypotheses, which amounts to overconstraining the space of possibilities. *This last form of forgetting is equivalent to a local preference bias* which chooses among competing hypotheses which ones to pursue.

This raises then a more specific question than the aforementioned ones, but which contributes to the same overall goal :

---

[2] Finding the global optimum at each step requires either to keep in memory all possible hypotheses, re-evaluate them and take the best one, or, short to this memory intensive method, to be able to re-construct any possible hypothesis with enough accuracy and keep track to the best one so far. This latter method is closed in spirit to Simulated Annealing or to Genetic Algorithms. In mathematics, this corresponds to ergodic systems.

3. *In which case an incremental learner can be order independent ? Or, in other words, which information can be safely forgotten without altering the result of learning whatever is the ordering of inputs ?*

It is this last question that this paper focuses on. It must be kept in mind that it is equivalent to the question : what local preference bias leads to a null global preference bias (i.e. to order independence) ?

In the following of the paper, **we will restrict ourselves to a simple concept learning model** in which the learner attempts to infer an unknown target concept f, chosen from a known concept class F of {0,1}-valued functions over an instance space X. This framework allows us, in the next section, to define a measure of the information gained by the learning system and of the effect of a local bias on this information. This measure naturally suggests an equivalence relation between local preference bias and additional instances, which is detailed in section 4. Then, in section 5, it becomes a relatively simple matter to answer question 3 above. The conclusion compares the framework adopted here with the emerging one of teachability and discusses the results obtained.

3   Information measure and local preference bias

In this section, we are interested in formalizing and quantifying the effect of a local preference bias on what is learned by the system. For this, we first define a characterization of the information maintained by a learner.

Let F be a concept class over the instance space X, and $f \in$ F be a target concept. The teacher has a collection of examples EX=$\{x_i, f(x_i)\}$ at his disposal, and makes a sequence $\mathbf{x}=x_1, x_2, \ldots, x_m, x_{m+1}, \ldots$ with $x_m \in$ EX for all m. The learner receives information about f incrementally via the label sequence $f(x_1), \ldots, f(x_m), f(x_{m+1}), \ldots$ For any $m \geq 1$, we define (with respect to $\mathbf{x}, f$) the mth *version space* :

$$F_m(\mathbf{x}, f) = \{\hat{f} \in F : \hat{f}(x_1) = f(x_1), \ldots, \hat{f}(x_m) = f(x_m)\}$$

The version space at time m is simply the class of all concepts in F consistent with the first m labels of f (with respect to $\mathbf{x}$). $F_m(\mathbf{x}, f)$ will serve as a **characterization**

of what is known to the learner at time m about the target concept f.

We know from Mitchell [10] that the version space can be economically represented and stored using the boundary sets S-set (set of the most general hypotheses that are more specific than the concepts in the version space), and G-set (set of the most specific hypotheses that are more general than the concepts in the version space)[3]. Each new example $(x_m, f(x_m))$ provides new information if it allows to reduce the version space by modifying, through the CE algorithm, either one of the boundary sets. Generally, the S-set and the G-set contain many elements, and in worst cases, they can grow exponentially over some sequences of examples [6].

A local preference bias is a choice strategy which, at any time m, discards parts of the current version space, generally in order to keep the boundary sets manageable. In this way, it reduces the version space and acts as if there had been some additional information that had allowed to constrain the space of hypotheses. The next section gives a closer look at this equivalence.

4   Bias and additional instances

We assume that the incremental learner maintains a version space of potential concepts by keeping the boundary sets. We assume further that the local preference bias, if any, acts by removing elements of the S-set and/or of the G-set, thus reducing the version space. Indeed, in so doing, it removes from the version space all concepts or hypotheses that are no longer more general than some element of the S-set and more specific than some element of the G-set. Besides, the resulting version space keeps its consistency since, in this operation, no element of the resulting S-set become more general than other elements of the S-set or of the G-set, and vice-versa, no element of the G-set can become more specific than other elements of the G-set or of the S-set.

----

[3] To be exact, a subset of the concept space can be represented by its S-set and G-set if and only if it is closed under the partial order of generality of the description language and bounded. This is the case for most learning tasks and concept representations and particularly when the description language consists in the set of all conjunctive expressions over a finite set of boolean features. See [7] for more details.

To sum up, we now have a learning system that forgets pieces of information during learning by discarding potential hypotheses, but at the same time is optimal since, in principle, by keeping the set of all the remaining hypotheses, it could select the best among them. In that way, we isolate the effect of forgetting without intermingling with non optimality effects.

We are interested in studying the action of a local preference bias (which has been shown to be equivalent to the forgetting of potential hypotheses) along all possible sequences of data. More specifically, we want to find what type of local bias (or forgetting strategy) leads to no ordering effects, that is for which all training sequences conduct to the same resulting state.
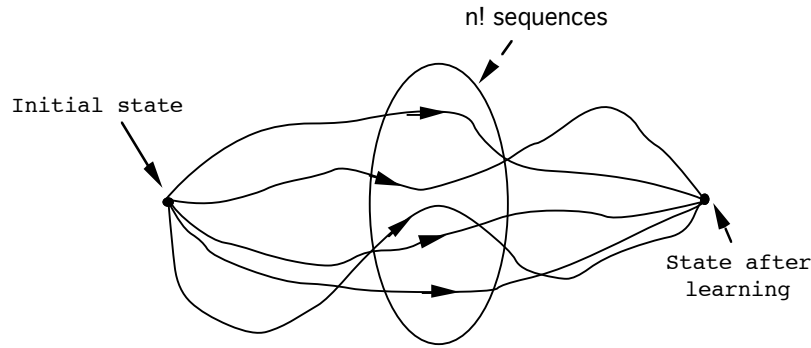
Fig. 2. For *n* instances, there are *n!* possible training sequences. We look for conditions under which all sequences would result in the same state.

What makes this problem difficult is that the action of the local preference bias depends on which state the system is in, and therefore depends on the training sequence followed. This implies in turn that all training sequences should be compared in order to find conditions on the local bias.

A very simple but very important idea will allow to circumvent this obstacle. It has three parts :

(i) forgetting hypotheses amounts to overconstrain the search space

(ii) extra instances to an order independent learning algorithm would result in constraining the search space

(iii) if an incremental learner using a local bias b_1 (leading to forgetting of hypotheses) could be made

equivalent to an order independent incremental learner using bias b_2, (leading to the consideration of extra instances) then, finding conditions on the local bias b_1 would be the same as finding conditions on the bias b_2, only this time irrelevant of the training sequence (since b_2 is used by an order independent learner).

(i) and *(ii)* allow to realize *(iii)* if it can be shown that the effect of the local bias b_1 is the same as the effect of additional instances given by an oracle or bias b_2 to an order independent learner. In other words, if it can be proved that any forgetting of hypothesis is equivalent to observing extra instances, then, conditions on b_1 will amount to conditions on the addition of fictitious instances to an order independent learning algorithm.

Order dependent learner using
local bias (forgetting)

Order independent algorithm
(e.g. the CE algorithm)
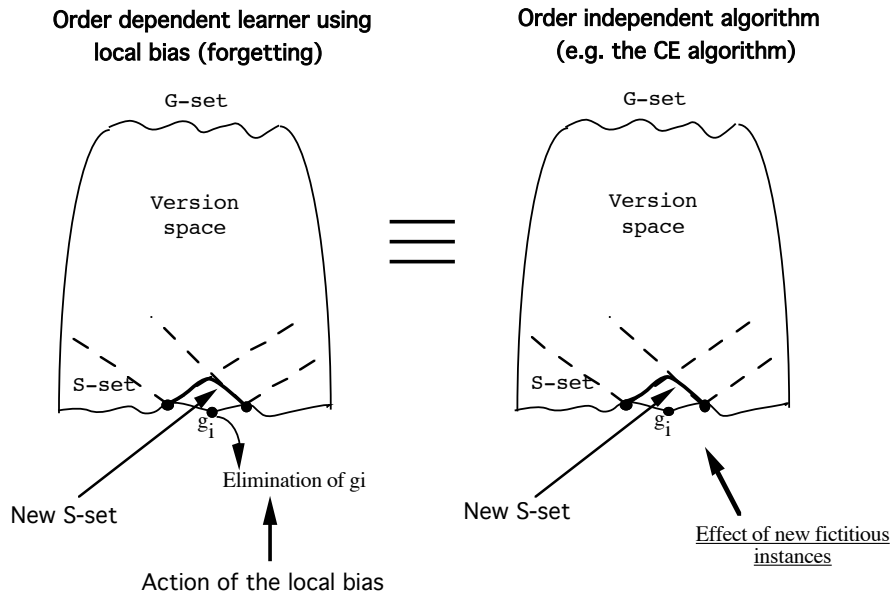
Fig. 3. The equivalence needed to allow the study of conditions on the local bias. The effect of the local bias (forgetting of some element of the S-set or G-set) is the same as the effect of additional instances made available to an order independent algorithm such as the CE algorithm.

Of this powerful idea, we will make a theorem.

Theorem 1 : *With each choice it makes, the local preference bias acts as if additional examples had been known to an order independent learner.*

Proof : (i) Case of the reduction of the S-set. For each element g_i of the S-set it is possible to find additional fictitious examples which, if considered by the CE algorithm, would lead to its elimination of the S-set. It suffices to take the positive instances covered by (or more specific than) all g_j such that ($g_j \in$ S-set and j≠i) and not covered by g_i or excluded by the G-set. As a result, the CE algorithm would not have to modify the G-set nor the g_j such that ($g_j \in$ S-set and j≠i) and it would generalize g_i just enough to cover the new instances. But since {g_j / ($g_j \in$ S-set and j≠i)} is the S-set of all past instances plus the new fictitious ones, g_i can only become more general than one or several g_j, and hence will be eliminated from the new S-set.

(ii) Case of the reduction of the G-set. In the same way, in order to eliminate an element g_i of the G-set through the CE algorithm, it suffices to provide the negative instances covered by g_i but not covered by the other elements of the G-set and by the S-set. As for (i) above, the CE algorithm would then specialize g_i just enough to exclude the negative instances, and this would result in an element of the G-set that would be more specific than others, hence eliminated. $\mathbb{P}$
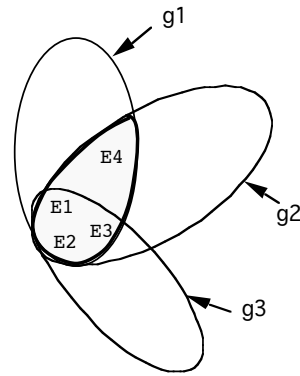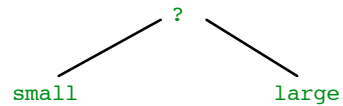

Fig. 4. In this figure, {g1,g2,g3} are assumed to be the S-set of the positive instances (E1,E2,E3). If E4 is observed, then neither g1 nor g2 need to be modified. In fact {g1,g2} is the S-set of all positive instances that belong to the gray area. g3 will need to be generalized just enough to cover E4, and this will make it more general than g1 and g2, hence it will be eliminated from the S-set.
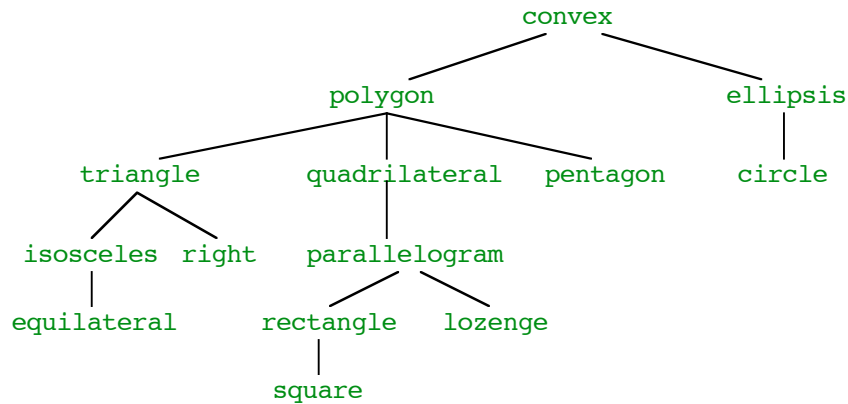
The following **example** will help to understand this.

Let us assume that we present positive and negative instances of scenes made of several objects each one described by a conjunction of attribute-values, and that we use three families of attributes, each one organized as a tree with increasing order of generality toward the root.
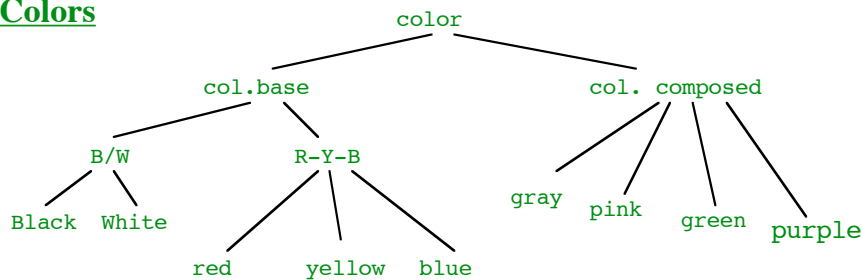
Sizes

```
              ?
          /       \
      small        large
```

Shapes

```
                        convex
                      /        \
                polygon         ellipsis
              /    |    \          |
       triangle quadrilateral pentagon   circle
        /    \        |
  isosceles right  parallelogram
      |            /       \
 equilateral  rectangle   lozenge
                  |
               square
```

Colors

```
                    color
                  /        \
            col.base        col. composed
            /      \        /   |    \    \
        B/W        R-Y-B  gray pink green  purple
        /  \      /  |  \
   Black White  red yellow blue
```

In the following, we show how an element of the S-set could be removed by observing more instances.

Let us suppose that we observe the following sequence of positive instances :

- E1:   +   {(large red square) & (small white lozenge)}

∅   <u>S-set-1</u> = {(large red square) & (small white lozenge)}

- E2:   +   {(large red parallelogram) & (small blue equilateral)}

∅   <u>S-set-2</u> = {(large red parallelogram) & (small col.-base polygon),

                  (? R-Y-B polygon) & (? col.-base parallelogram)}

- E3: + {(large yellow right) & (small blue rectangle)}

∅   <u>S-set-3</u> = {(large R-Y-B polygon) & (small col.-base polygon),   ◆ g 1

                  (? R-Y-B parallelogram) & (? col.-base polygon),   ◆ g 2

                  (? R-Y-B polygon) & (? col.-base parallelogram)}   ◆ g 3

Let us assume that at this point the local bias for some reasons decides to discard g3 from the S-set. This would result in

S-set-4 = {(large R-Y-B polygon) & (small col.-base polygon),

(? R-Y-B parallelogram) & (? col.-base polygon)}.

The very same S-set would be obtained if the learning algorithm was the CE algorithm that after `E1`, `E2` and `E3` observed a new instance covered by g1 and by g2 and not by g3 such as :

```
-E4: +      {(large yellow parallelogram) & (small black pentagon)}
```

5   Bias and order independence

What we have seen so far is that a local preference bias (forgetting hypotheses) can be made equivalent to another bias that would throw in chosen extra instances for the learner to observe. Thanks to this we can now tackle the main topic of this paper, namely what kind of local preference bias a learner can implement so as to stay order independent. Indeed, instead of studying a strategy of forgetting of hypotheses that depend on the current version space, and therefore on the past history, we now study addition of extra fictitious instances to the CE algorithm that is order independent. In other words, we are now in a position to specify conditions on the local preference bias by stating to which extra instances it should amount to.

We assume that the teacher has a set of n examples EX, and draws a sequence $\mathbf{x}$ of these according to her requirements.

Furthermore, we assume order independence, i.e. :

(1) $\forall \mathbf{x}, F_n^{LB}(\mathbf{x}, f) = VS_{wb}$,   where

         VS_{wb} is constant.

(We use the notation F_n^{LB} to differentiate a learner implementing a local bias (LB) from one that does not and only implements the CE algorithm noted $F_n(\mathbf{x}, f)$ in section 3. VS_{wb} means the version space obtained with bias).

Theorem   2 : *An incremental learner implementing a local preference bias is order independent for a collection EX of instances if the action of this bias is equivalent for all possible sequences $\mathbf{x}$ of elements of EX to the supply of the same set of additional instances.*

Proof : It follows immediately from theorem 1. $\mathbb{P}$

Now, we want to enlarge theorem 2, and give conditions upon the set of fictitious examples that the local bias acting as an oracle can provide to the learner so that the learner be order independent and reach VS_{wb} possibly different from VS_{nb}.

The CE algorithm without bias would reach the state VS_{nb} (Version Space with no bias). If $VS_{wb} \neq VS_{nb}$, then it follows that extra instances should be provided to the CE algorithm so that it reaches VS_{wb}. Theorem 3 states which ones.
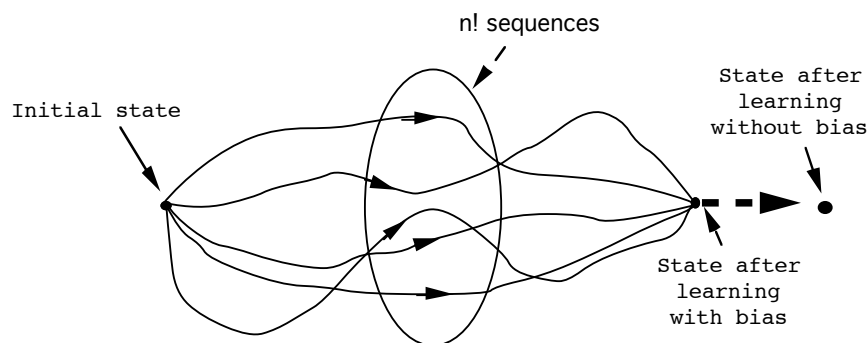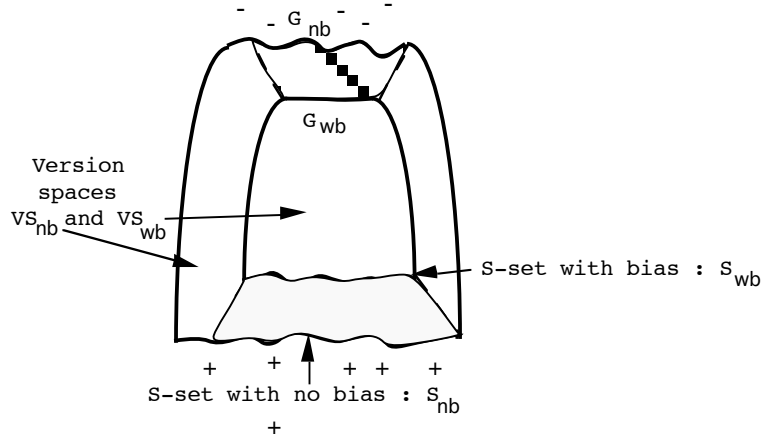
Fig. 5. Which fictitious instances should the oracle corresponding to the local bias provide to the CE algorithm so that it get to the version space VS_{wb} ? Answering this question amounts to give conditions on the local bias that would conduct a learner using it to VS_{wb} irrespective of the training sequence followed. The answer is that fictitious positive instances should be drawn from the gray array, whereas fictitious negative instances should be drawn from the stripped array. The '+' and '-' figure the real instances that conduct the CE algorithm to the VS_{nb} version space.

Let S_{nb} and G_{nb} be respectively the S-set and the G-set that the CE algorithm would obtain from the collection of instances EX, and let S_{wb} and G_{wb} be respectively the S-set and the G-set of VS_{wb} in (1) (the version space obtained on any sequence **x** of EX by the learner implementing the local preference bias).

Theorem 3 : *For an incremental learner implementing a local preference bias to be order independent for EX leading to the version space C, it is necessary that the action of this bias be equivalent to the supply of :*

- a <u>set of fictitious positive instances</u> such that they together with the real positive instances are covered and bounded by S_{wb}, and,

- a <u>set of fictitious negative instances</u> such that they together with the real negative ones are excluded and bounded by G_{wb}.

Proof : Self-evident. ℙ

The next theorem is the application of theorem 3 to the case where $VS_{wb} = F_n(EX,f) = VS_{nb}$, that is the case where the local preference bias leads to the null global preference bias, i.e. has no effect. Such a local bias can be seen as eliminating options judiciously since the result obtained after any sequence **x** of EX is the same as what the CE algorithm would get on EX. In this case S_{nb} and S_{wb} are one and the same as are G_{nb} and G_{wb}.

Theorem 4 : *For an incremental learner implementing a local preference bias leading to the same result as an incremental learner without a local bias, it is necessary that the action of this bias be equivalent to the supply of :*

- a set of positive instances such that each one is covered by elements of S_{nb}, and ,

- a set of negative instances such that each one is covering all elements of G_{nb}.

It is as if this local preference bias knew "in advance" the collection EX of instances, and eliminated elements of the S-set and of the G-set judiciously. This leads to the final theorem.

Theorem 5 : *It is not possible for a local deterministic preference bias to lead to a null global preference bias for any arbitrary collection EX of examples.*

Proof : Indeed, at each step, the action of the local bias can only depend on the past training instances and the current state. In order to lead to order independence it would have to be equivalent to an oracle that provides instances drawn from an array of the instance space that can be defined only with respect to all the training instances. This would mean that the learner, through its preference bias, was always perfectly informed in advance on the collection EX of examples held by the teacher. ℙ

6 Conclusion

In this research, we are interested in the following **general question** : *given a collection of examples (or data in general), how can a teacher, a priori, best put them in sequence so that the learner, a deterministic incremental learning system that does not ask questions during learning, can acquire some target concept (or knowledge)?*

This question, that corresponds to situations where the teacher does not have the choice of the examples and can not interpret the progress made by the student until the end of the learning period, leads to the study of incremental learning per se, independently of any particular system. The solution to this general interrogation could be of some use to several realistic settings where a teacher has an incremental learning system and a collection of data, or when data arrive sequentially but time allows to keep them in small buffers that can be ordered before being processed.

This framework is **to be compared with** the recent surge of interest for "teaching strategies" that allow to optimally teach a concept to a learner, thus providing lower bounds on learnability complexity [5,17]. The difference with the former framework is that in one case the teacher can only play on the order of the sequence of inputs, whether in the other case, the teacher chooses the most informative ideal examples but does not look for the best order (there are some exceptions such as [13]).

This paper has outlined some first results concerning order sensitivity in supervised conceptual incremental learning. The most important ones are :

(i) that order dependence is due non-optimality and/or to forgetting of possibilities corresponding to a local preference bias that heuristically selects the most promising hypotheses,

(ii) that this bias can be seen as the result of additional instances given to the learner (i.e. prior knowledge built into the system),

(iii) that (ii) allows to replace the difficult problem of determining the action of the local bias along different sequences of instances by a problem of addition (which is commutative, i.e. order independent) of instances to an order independent learner, which leads to

(iv) that there are strong contingencies for an incremental learner to be order independent on some collections of instances (either the corresponding prior knowledge is well-tailored to the future potential collections of inputs, or there is no prior knowledge, thus no reduction of storage and computational complexity).

In this study, we have given sufficient conditions only on the fictitious examples that should be provided by an oracle equivalent to a local preference bias. It would be nice to obtain necessary conditions that state which instances should necessarily be given by the oracle to get order independence. We are currently working on this question.

It should be clear that the problem of noisy data is of no concern to us here. We characterize through the version space what is known to the learner irrespective of the quality of the data. Noise would become an important issue if it was dependent on the ordering of the data (e.g. a sensor equipment that would degrade with time).

Issues for future research include : are these results extensible to more general learning situations (e.g. unsupervised) ? given a local preference bias, how to determine a good sequence ordering so as to best guide the system towards the target knowledge ?

References

1. **Cornuéjols** A. (1989) : "*An Exploration into Incremental Learning : the INFLUENCE System*", in Proc.of the 6th Intl. Conf. on Machine Learning, Ithaca, June 29- July 1, 1989, pp.383-386.

2. **Daley & Smith** (1986) : "*On the Complexity of Inductive Inference*". Information and Control, 69, pp.12-40, 1986.

3. **Doyle** J. (1979) : "*A truth maintenance system*". Artificial Intelligence, 12, 231-272, 1979.

4. **Fisher, Xu & Zard** (1992) : "*Ordering Effects in COBWEB and an Order-Independent Method*". To appear in Proc. of the 9th Int. Conf. on Machine Learning, Aberdeen, June 29-July 1st, 1992.

5. **Goldman & Kearns** (1991) : "*On the Complexity of Teaching*". Proc. of COLT'91, Santa Cruz, Aug. 5-7 1991, pp.303-314.

6. **Haussler** D. (1988) : "*Quantifying inductive bias: AI learning algorithms and Valiant's learning framework*". Artificial Intelligence, 36, 177-222.

7. **Hirsh** H. (1990) : "*Incremental Version-Space Merging*". Proc. of the 7th Int. Conf. on Machine Learning. Univ. of Austin, Texas, June 21-23, 1990, pp.330-338.

8. **Ling** (1991) : "*Inductive Learning from Good Examples*". In Proc. of the 12th Int. Joint Conf. on Artif. Intel. (IJCAI-91), pp.751-756.

9. **MacGregor** J. (1988) : "*The Effects of Order on Learning Classifications by Example: Heuristics for Finding the Optimal Order*", Artificial Intelligence, vol.34, pp.361-370, 1988.

10. **Mitchell** T. (1978) : *Version Spaces : an Approach to Concept Learning*. PhD Thesis, Stanford, December 1978.

11. **Mitchell** T. (1982) : "*Generalization as Search*", Artificial Intelligence, vol.18, pp.203-226, 1982.

12. **Natarajan** B. (1991) : *Machine Learning. A Theoretical Approach*. Morgan Kaufmann, 1991.

13. **Pearl** J. (1988) : *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

14. **Porat & Feldman** (1991) : "*Learning Automata from Ordered Examples*". Machine Learning,7, pp.109-138, 1991.

15. **Rivest & Sloan** (1988) : "*Learning Complicated Concepts Reliably and Usefully*". In Proc. of COLT'88 (Workshop on Computational Learning Theory), Cambridge 1988, pp.69-79.

16. **Shinohara & Miyano** (1990) : "*Teachability in Computational Learning*", in Proc. of the Workshop on Algorithmic Learning Theory, 1990, pp.247-255.

17. **Salzberg, Delcher, Heath & Kasif** (1991) : "*Learning with a helpful teacher*". Proc. of the IJCAI-91, pp.705-711.

18. **Utgoff** P. (1989) : "*Incremental Induction of Decision Trees*". Machine Learning, 4,161-186,(1989).

19. **Van Lehn** (1987) : "*Learning one subprocedure per lesson*". Artificial Intelligence, 31 (1), pp.1-40, january 1987.

20. **Winston** (1970) : "*Learning structural descriptions from examples*", AI-TR-231, MIT, Artificial Intelligence Laboratory, Cambridge, MA, 1970.