Preference-Based Batch and Sequential Teaching: Towards a Unified View of Models

Farnam Mansouri† Yuxin Chen‡ Ara Vartanian∗ Xiaojin Zhu ‡ Adish Singla†

†Max Planck Institute for Software Systems (MPI-SWS), {mfarnam, adishs}@mpi-sws.org.
‡University of Chicago, chenyuxin@uchicago.edu.
∗University of Wisconsin-Madison, {aravart, jerryzhu}@cs.wisc.edu

Abstract

Algorithmic machine teaching studies the interaction between a teacher and a learner where the teacher selects labeled examples aiming at teaching a target hypothesis. In a quest to lower teaching complexity and to achieve more natural teacher-learner interactions, several teaching models and complexity measures have been proposed for both the batch settings (e.g., worst-case, recursive, preference-based, and non-clashing models) as well as the sequential settings (e.g., local preference-based model). To better understand the connections between these different batch and sequential models, we develop a novel framework which captures the teaching process via preference functions Σ. In our framework, each function σ ∈ Σ induces a teacher-learner pair with teaching complexity as TD(σ). We show that the above-mentioned teaching models are equivalent to specific types/families of preference functions in our framework. This equivalence, in turn, allows us to study the differences between two important teaching models, namely σ functions inducing the strongest batch (i.e., non-clashing) model and σ functions inducing a weak sequential (i.e., local preference-based) model. Finally, we identify preference functions inducing a novel family of sequential models with teaching complexity linear in the VC dimension of the hypothesis class: this is in contrast to the best known complexity result for the batch models which is quadratic in the VC dimension.

1 Introduction

Algorithmic machine teaching studies the interaction between a teacher and a learner where the teacher’s goal is to find an optimal training sequence to steer the learner towards a target hypothesis [GK95, ZLHZ11, Zhu13, SBB+14, Zhu15, ZSZR18]. An important quantity of interest is the teaching dimension (TD) of the hypothesis class, representing the worst-case number of examples needed to teach any hypothesis in a given class. Given that the teaching complexity depends on what assumptions are made about teacher-learner interactions, different teaching models lead to different notions of teaching dimension. In the past two decades, several such teaching models have been proposed, primarily driven by the motivation to lower teaching complexity and to find models for which the teaching complexity has better connections with learning complexity measured by Vapnik–Chervonenkis dimension (VCD) [VC71] of the class.

Most of the well-studied teaching models are for the batch setting (e.g., worst-case [GK95, Kuh99], recursive [ZLHZ08, ZLHZ11, DFSZ14], preference-based [GRSZ17], and non-clashing [KSZ19] models). In these batch models, the teacher first provides a set of examples to the learner and then the learner outputs a hypothesis. In a quest to achieve more natural teacher-learner interactions and enable richer applications, various different models have been proposed for the sequential setting (e.g., local preference-based model for version space learners [CSMA+18], models for gradient

learners [LDH+17] [LDL+18] [KDCST19], models inspired by control theory [Zhu18] [LZZ19], and models for human-centred applications that require adaptivity [SBB+13] [HCMA+19].

In this paper, we seek to gain a deeper understanding of how different teaching models relate to each other. To this end, we develop a novel teaching framework which captures the teaching process via preference functions $\Sigma$. Here, a preference function $\sigma \in \Sigma$ models how a learner navigates in the version space as it receives teaching examples (see §2 for formal definition); in turn, each function $\sigma$ induces a teacher-learner pair with teaching dimension $TD(\sigma)$ (see §3). We highlight some of the key results below:

- We show that the well-studied teaching models in batch setting corresponds to specific families of $\sigma$ functions in our framework (see §4 and Table 1).
- We study the differences in the family of $\sigma$ functions inducing the strongest batch model [KSZ19] and functions inducing a weak sequential model [CSMA+18] (§5.2) (also, see the relationship between $\Sigma_{gvs}$ and $\Sigma_{local}$ in Figure 1).
- We identify preference functions inducing a novel family of sequential models with teaching complexity linear in the VCD of the hypothesis class. We provide a constructive procedure to find such $\sigma$ functions with low teaching complexity (§5.3).

Our key findings are highlighted in Figure 1 and Table 1. Here, Figure 1 illustrates the relationship between different families of preference functions that we introduce, and Table 1 summarizes the key complexity results we obtain for different families. Our unified view of the existing teaching models in turn opens up several intriguing new directions such as (i) using our constructive procedures to design preference functions for addressing open questions of whether RTD/NCTD is linear in VCD, and (ii) understanding the notion of collusion-free teaching in sequential models. We discuss these directions further in §6.

<table>
<thead>
<tr>
<th>Families</th>
<th>$\Sigma_{const}$</th>
<th>$\Sigma_{global}$</th>
<th>$\Sigma_{gvs}$</th>
<th>$\Sigma_{local}$</th>
<th>$\Sigma_{lvs}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduction</td>
<td>TD</td>
<td>RTD/PBTD</td>
<td>NCTD</td>
<td>Local-PBTD</td>
<td>–</td>
</tr>
<tr>
<td>Complexity Results</td>
<td>O(VCD)</td>
<td>O(VCD)</td>
<td>O(VCD)</td>
<td>O(VCD)</td>
<td>O(VCD)</td>
</tr>
<tr>
<td>[GK95]</td>
<td>[ZLHZ11]</td>
<td>[GRSZ17]</td>
<td>[HHLW17]</td>
<td>[KSZ19]</td>
<td>[CSMA+18]</td>
</tr>
</tbody>
</table>

Table 1: Overview of our main results: Reduction to existing models and teaching complexity.

2 The Teaching Model

The teaching domain. Let $\mathcal{X}$, $\mathcal{Y}$ be a ground set of unlabeled instances and the set of labels. Let $\mathcal{H}$ be a finite class of hypotheses; each element $h \in \mathcal{H}$ is a function $h : \mathcal{X} \rightarrow \mathcal{Y}$. Here, we only consider boolean functions and hence $\mathcal{Y} = \{0, 1\}$. In our model, $\mathcal{X}$, $\mathcal{H}$, and $\mathcal{Y}$ are known to both the teacher and the learner. There is a target hypothesis $h^* \in \mathcal{H}$ that is known to the teacher, but not the learner. Let $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ be the ground set of labeled examples. Each element $z = (x_z, y_z) \in \mathcal{Z}$ represents a labeled example where the label is given by the target hypothesis $h^*$, i.e., $y_z = h^*(x_z)$. For any $Z \subseteq \mathcal{Z}$, the version space induced by $Z$ is the subset of hypotheses $\mathcal{H}(Z) \subseteq \mathcal{H}$ that are consistent with the labels of all the examples, i.e., $\mathcal{H}(Z) := \{h : h \in \mathcal{H} \text{ and } \forall z \in Z, h(x_z) = y_z\}$.

Learner’s preference function. We consider a generic model of the learner that captures our assumptions about how the learner adapts her hypothesis based on the labeled examples received from the teacher. A key ingredient of this model is the learner’s “preference function” over the hypotheses. The learner, based on the information encoded in the inputs of preference function—which include the current hypothesis and the current version space—will choose one hypothesis in $\mathcal{H}$. Our model of the learner strictly generalizes the local preference-based model considered in [CSMA+18], where the learner’s preference was only encoded by her current hypothesis. Formally, we consider preference functions of the form $\sigma : \mathcal{H} \times 2^\mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$. For any two hypotheses $h', h''$, we say that the learner prefers $h'$ to $h''$ based on the current hypothesis $h$ and version space $H \subseteq \mathcal{H}$, i.e., $\sigma(h'; H, h) < \sigma(h''; H, h)$. If $\sigma(h'; H, h) = \sigma(h''; H, h)$, then the learner could pick either one of these two.
An important consideration when designing teaching models is to ensure that the teacher and the learner use the same preference function. This assumption of shared knowledge of the preference function is also considered in existing teaching models for both the batch settings (e.g., as in [ZLHZ11, GRSZ17]) and the sequential settings [CSMA 18].

### Protocol 1

**Interaction protocol between the teacher and the learner**

1. learner’s initial version space is $H_0 = \mathcal{H}$ and learner starts from an initial hypothesis $h_0 \in \mathcal{H}$
2. for $t = 1, 2, 3, \ldots$ do
   3. learner receives $z_t = (x_t, y_t)$; updates $H_t = H_{t-1} \cap \mathcal{H}(\{z_t\})$; picks $h_t$ per Eq. (2.1);
   4. teacher receives $h_t$ as feedback from the learner;
   5. if $h_t = h^*$ then teaching process terminates

### 3 The Complexity of Teaching

#### 3.1 Teaching Dimension for the Sequential Setting

Our objective is to design teaching algorithms that can steer the learner towards the target hypothesis $h^*$ by providing a sequence of labeled examples. The learner starts with an initial hypothesis $h_0 \in \mathcal{H}$ before receiving any labeled examples from the teacher. At time step $t$, the teacher selects a labeled example $z_t \in \mathcal{X} \times \mathcal{Y}$, and the learner makes a transition from the current hypothesis to the next hypothesis. Let us denote the labeled examples received by the learner up to (and including) time step $t$ via $Z_t$. Further, we denote the learner’s version space at time step $t$ as $H_t = \mathcal{H}(Z_t)$, and the learner’s hypothesis before receiving $z_t$ as $h_{t-1}$. The learner picks the next hypothesis based on the current hypothesis $h_{t-1}$, version space $H_t$, and preference function $\sigma$:

$$h_t \in \arg \min_{h' \in H_t} \sigma(h'; H_t, h_{t-1}). \quad (2.1)$$

Upon updating the hypothesis $h_t$, the learner sends $h_t$ as feedback to the teacher. Teaching finishes here if the learner’s updated hypothesis $h_t = h^*$. We summarize the interaction in Protocol 1.

### 3.2 $\Sigma$-TD: Teaching Dimension for a Family of Preference Functions

An important consideration when designing teaching models is to ensure that the teacher and the learner are “collusion-free”, i.e., they are not allowed to collude or use some “coding-trick” to achieve arbitrarily low teaching complexity. A well-accepted notion of collusion-freeness in the batch setting is one proposed by [GM96] (also see [AK97, OS99, KSZ19]). Intuitively, it captures the idea that a learner conjecturing hypothesis $h$ will not change its mind when given additional information consistent with $h$. In comparison to batch models, sequential models are much less understood, there is no established notion of collusion-freeness, and their relations to batch models is not clear.

---

1 It is important to note that in our teaching model, the teacher and the learner use the same preference function. This assumption of shared knowledge of the preference function is also considered in existing teaching models for both the batch settings (e.g., as in [ZLHZ11, GRSZ17]) and the sequential settings [CSMA 18].
Collusion-free preference functions. We now introduce a class of preference functions, which we call collusion-free preference functions. We extend the collusion-free definition of [GM96] to the sequential setting, by assuming that when \( \hat{h} \) is the only hypothesis in the most preferred set defined by \( \sigma \), then the learner will not change its mind if given additional information consistent with \( \hat{h} \). We formalize this notion in the following definition:

**Definition 1 (Collusion-free preference)** Assume that the teacher has proposed a sequence of teaching examples \( Z_t \) at time \( t \), and the learner’s current hypothesis is \( h_{t-1} \). We call a preference function collusion-free, if for any \( t > 0 \), \( \arg \min_{h' \in H_t} \sigma(h'; H_t, h_{t-1}) = \{ h \} \) implies that \( \arg \min_{h' \in H_t \cap \mathcal{H}(S)} \sigma(h'; H_t \cap \mathcal{H}(S), h) = \{ h \} \) for any set of examples \( S \) provided by the adversary which are consistent with \( h \).

In this paper, we focus on preference functions that are collusion-free. In particular, we use \( \Sigma_{\text{CF}} \) to denote the family of preference functions that induce collusion-free teacher-learner pairs:

\[
\Sigma_{\text{CF}} = \{ \sigma | \sigma \text{ is collusion-free} \}.
\]

Teaching dimension \( \Sigma\cdot\text{TD} \). In the subsequent sections, we will investigate several subclasses of the collusion-free preference functions. For a family of preference functions \( \Sigma \subseteq \Sigma_{\text{CF}} \), we define the teaching dimension w.r.t the family \( \Sigma \) as the teaching dimension w.r.t. the best \( \sigma \) in that family:

\[
\Sigma\cdot\text{TD}_{X,\mathcal{H},h_0} = \min_{\sigma \in \Sigma} \text{TD}_{X,\mathcal{H},h_0}(\sigma)
\]

### 4 Preference-based Batch Models

#### 4.1 Families of Preference Functions

We consider three families of preference functions which do not depend on the learner’s current hypothesis. The first one is the family of uniform preference functions, denoted by \( \Sigma_{\text{const}} \), which corresponds to constant preference functions:

\[
\Sigma_{\text{const}} = \{ \sigma \in \Sigma_{\text{CF}} | \exists c \in \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = c \}
\]

The second family, denoted by \( \Sigma_{\text{global}} \), corresponds to the preference functions that do not depend on the learner’s current hypothesis and version space. In other words, the preference functions capture some global preference ordering of the hypotheses:

\[
\Sigma_{\text{global}} = \{ \sigma \in \Sigma_{\text{CF}} | \exists g : \mathcal{H} \to \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = g(h') \}
\]

The third family, denoted by \( \Sigma_{\text{gvs}} \), corresponds to the preference functions that depend on the learner’s version space, but do not depend on the learner’s current hypothesis:

\[
\Sigma_{\text{gvs}} = \{ \sigma \in \Sigma_{\text{CF}} | \exists g : \mathcal{H} \times 2^\mathcal{H} \to \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = g(h', H) \}
\]

Figure 2 illustrates the relationship between these preference families.

#### 4.2 Complexity Results

We first provide several definitions, including the formal definition of VC dimension as well as several classical notions of teaching dimension.

**Definition 2 (Vapnik–Chervonenkis dimension)** The VC dimension for \( H \subseteq \mathcal{H} \) w.r.t. a fixed set of unlabeled instances \( X \subseteq X \), denoted by \( \text{VCD}(H, X) \), is the cardinality of the largest set of points \( X' \subseteq X \) that are “shattered”. Formally, let \( H|_X = \{(h(x_1), ..., h(x_n)) | \forall h \in H \} \) denote all possible patterns of \( H \) on \( X \). Then \( \text{VCD}(H, X) = \max |X'|, \text{ s.t. } X' \subseteq X \) and \( |H|_X = 2^{|X'|} \).

**Definition 3 (Teaching dimension)** For any hypothesis \( h \in \mathcal{H} \), we call a set of teaching instances \( T(h) \subseteq X \) a teaching set of \( h \), if it can uniquely identify \( h \in \mathcal{H} \). The teaching dimension for \( \mathcal{H} \), denoted by \( \text{TD}(\mathcal{H}) \), is the maximum size of the minimum teaching set for any \( h \in \mathcal{H} \):

\[
\text{TD}(\mathcal{H}) = \max_{h \in \mathcal{H}} \min_{T(h)} |T(h)|.
\]

\(^2\)Note that in the classical definition of VCD, only the first argument of a hypothesis class \( H \) is present; the second argument of \( X \) is omitted and is by default the whole ground set of unlabeled instances \( X \).
As noted by [ZLHZ08], the teaching dimension of [GK95] does not always capture the intuitive idea of cooperation between teacher and learner. The authors then introduced a model of cooperative teaching that resulted in the complexity notion of recursive teaching dimension, as defined below.

**Definition 4 (Recursive teaching dimension [ZLHZ08 DSZ10])** The recursive teaching dimension (RTD) of $H$, denoted by $\text{RTD}(H)$, is the smallest number $t$ where one can order all the hypothesis in $H$ as an ordered sequence $(h_1, \ldots, h_t)$, such that every hypothesis $h_i$, $i < |H|$, has a teaching set of size no more than $t$ to be distinguished from the later hypotheses in the sequence.

Note that RTD is equivalent to preference-based teaching dimension (PBT) [GRSZ17] for finite hypothesis classes as we consider in this paper.

In a recent work of [KSZ19], a new notion of teaching complexity, called non-clashing teaching dimension or NCTD, was introduced (see definition below). Importantly, NCTD is the optimal teaching complexity among teaching models in the batch setting that satisfy the collusion-free property of [GM96].

**Definition 5 (Non-clashing teaching dimension [KSZ19])** Let $H$ be a hypothesis class and $T : H \to 2^X$ be a “teacher mapping” on $H$, i.e., mapping a given hypothesis to a teaching set. We say that $T$ is non-clashing on $H$ iff there are no two distinct $h, h' \in H$ such that both $T(h)$ is consistent with $h'$ and $T(h')$ is consistent with $h$. The No-Clash Teaching Dimension of $H$, denoted by $\text{NCTD}(H)$, is defined as $\text{NCTD}(H) = \min_{T \text{ is non-clashing}} \{\max_{h \in H} |T(h)|\}.

We show in the following, that the teaching dimension $\Sigma$-TD in Eq. (3.2) unifies the above classic definitions of TD’s for batch models.

**Theorem 1 (Reduction to classic notion of TD’s)** Fix $X, H, h_0$. The teaching complexity for the three families reduces to the classic notion of teaching dimensions:

1. $\Sigma_{\text{const}} \cdot \text{TD}_{X, H, h_0} = \text{TD}(H)$
2. $\Sigma_{\text{global}} \cdot \text{TD}_{X, H, h_0} = \text{RTD}(H) = O(\text{VCD}(H, X)^2)$
3. $\Sigma_{\text{gvs}} \cdot \text{TD}_{X, H, h_0} = \text{NCTD}(H) = O(\text{VCD}(H, X)^2)$

Our teaching model strictly generalizes the local-preference based model of [CSMA+18], which reduces to the “worst-case” model when $\sigma \in \Sigma_{\text{const}}$ (corresponding to TD) [GK95] and the global “preference-based” model when $\sigma \in \Sigma_{\text{global}}$. Hence we get $\Sigma_{\text{const}} \cdot \text{TD}_{X, H, h_0} = \text{TD}(H)$ and $\Sigma_{\text{global}} \cdot \text{TD}_{X, H, h_0} = \text{RTD}(H)$. To establish the equivalence between $\Sigma_{\text{gvs}} \cdot \text{TD}$ and NCTD, it suffices to show that for any $H$, it holds (i) $\Sigma_{\text{gvs}} \cdot \text{TD}_{X, H, h_0} \geq \text{NCTD}(H)$, and (ii) $\Sigma_{\text{gvs}} \cdot \text{TD}_{X, H, h_0} \leq \text{NCTD}(H)$. The full proof for the theorem is provided in Appendix 2 of the supplementary.

In Table 2, we consider the well known Warmuth hypothesis class [DSZ14], where $\Sigma_{\text{const}} \cdot \text{TD} = \text{TD} = 3$, $\Sigma_{\text{gvs}} \cdot \text{TD} = \text{RTD} = 3$, and $\Sigma_{\text{gvs}} \cdot \text{TD} = \text{NCTD} = 2$. Table 2b and Table 2d show preference functions $\sigma \in \Sigma_{\text{const}}$, $\sigma \in \Sigma_{\text{global}}$, and $\sigma \in \Sigma_{\text{gvs}}$ that achieve the minima in Eq. (3.2). Table 2c shows the teaching sequences achieving these teaching dimensions under different preference functions. In Appendix A.1 of the supplementary, we provide another hypothesis class where $\Sigma_{\text{const}} \cdot \text{TD} = 3$, $\Sigma_{\text{global}} \cdot \text{TD} = 2$ and $\Sigma_{\text{gvs}} \cdot \text{TD} = 1$.

5 Preference-based Sequential Models

5.1 Families of Preference Functions

In this section, we investigate two families of preference functions that depend on the learner’s current hypothesis $h_t$. The first one is the family of local preference-based functions [CSMA+18], denoted by $\Sigma_{\text{local}}$, which corresponds to preference functions that does not depend on the learner’s version space:

$\Sigma_{\text{local}} = \{\sigma \in \Sigma_{\text{CF}} | \exists g : H \times H \to \mathbb{R}, \text{s.t. } \forall h', H, h, \sigma(h', H, h) = g(h', h)\}$

The second family, denoted by $\Sigma_{\text{gvs}}$, corresponds to the preference functions that depend on all three arguments of $\sigma(h, h', H)$. The dependence of $\sigma$ on the learner’s (local) current hypothesis and the version space renders a powerful family of preference functions:

$\Sigma_{\text{gvs}} = \{\sigma \in \Sigma_{\text{CF}} | \exists g : H \times H \times \mathcal{H} \to \mathbb{R}, \text{s.t. } \forall h', H, h, \sigma(h', H, h) = g(h', H, h)\}$
We will use \( X \) to denote a compact-distinguishable set on \( H \) to denote a compact-distinguishable set on \( H \).

5.2 Comparing \( \Sigma_{\text{gvs}} \cdot \text{TD} \) and \( \Sigma_{\text{local}} \cdot \text{TD} \)

In the following, we show that substantial differences arise as we transition from \( \sigma \) functions inducing the strongest batch (i.e., non-clashing) model to \( \sigma \) functions inducing a weak sequential (i.e., local preference-based) model. We provide the full proof of Theorem 2 together with the construction procedure of examples satisfying these conditions in Appendix C of the supplementary.

**Theorem 2** Neither of the families \( \Sigma_{\text{gvs}} \) and \( \Sigma_{\text{local}} \) dominates the other. Specifically,

1. \( \Sigma_{\text{gvs}} \cap \Sigma_{\text{local}} = \Sigma_{\text{global}} \)
2. There exist \( \mathcal{H}, \mathcal{X} \), where \( \forall h_0 \in \mathcal{H}, \Sigma_{\text{local}} \cdot \text{TD}_{X, \mathcal{H}, h_0} > \Sigma_{\text{gvs}} \cdot \text{TD}_{X, \mathcal{H}, h_0} \)
3. There exist \( \mathcal{H}, \mathcal{X} \), where \( \forall h_0 \in \mathcal{H}, \Sigma_{\text{local}} \cdot \text{TD}_{X, \mathcal{H}, h_0} < \Sigma_{\text{gvs}} \cdot \text{TD}_{X, \mathcal{H}, h_0} \)

5.3 Complexity Results

We now connect the teaching complexity of the sequential models with the VC dimension.

**Theorem 3** \( \Sigma_{\text{local}} \cdot \text{TD}_{X, \mathcal{H}, h_0} = O(VCD(\mathcal{H}, \mathcal{X})^2) \), and \( \Sigma_{\text{gvs}} \cdot \text{TD}_{X, \mathcal{H}, h_0} = O(VCD(\mathcal{H}, \mathcal{X})) \).

To establish the proof, we first introduce a few useful definitions.

**Definition 6** (Compact-Distinguishable Set) Fix \( \mathcal{H} \subseteq \mathcal{H} \) and \( X \subseteq \mathcal{X} \), where \( X \) is given by \( X = \{x_1, \ldots, x_n\} \). Let \( H[X] = \{h(x_1), \ldots, h(x_n)\} \mid \forall h \in \mathcal{H} \} \) denote all possible patterns of \( H \) on \( X \). Then, we say that \( X \) is compact-distinguishable on \( H \), if \( |H[X]| = |H| \) and \( \forall X' \subset X, |H[X']| < |H| \). We will use \( \Psi_H \) to denote a compact-distinguishable set on \( H \).

---

\(^{3}\) We refer the reader to the original paper \([KSZ19]\) for a more formal description of “teacher mapping”.

\(^{4}\) The Warmuth hypothesis class is the smallest concept class for which RTD exceeds VCD.
In words, one can uniquely identify any hypothesis in $H$ with a (sub)set of examples from $\Psi_H$\cite{DFSZ14}. Furthermore, our definition of compact-distinguishable implies that there are no “redundant” examples in $\Psi_H$: (i) it does not contain any pair of distinct examples $x, x'$ such that $(\forall h \in H : h(x) = h(x'))$ or $(\forall h \in H : h(x) \neq h(x'))$; and (ii) it does not contain any $x$ such that $(\forall h \in H : h(x) = 1)$ or $(\forall h \in H : h(x) = 0)$.

The proof of the theorem relies on the following key lemma.

Lemma 4 Consider a subset $H \subseteq \mathcal{H}$ and any compact-distinguishable set $\Psi_H = \{x_1, \ldots, x_{|\Psi_H|}\}$. Fix any hypothesis $h_H \in H$. Let $d = \text{VCD}(H, \Psi_H)$ be the VC dimension of $H$ on $\Psi_H$. If $d \geq 1$, we can divide $H$ into $m = |\Psi_H| + 1$ separate hypothesis classes $\{H^1, \ldots, H^m\}$, such that

(i) $\forall j \in [m]$, there exists a compact-distinguishable set $\Psi_{H^j}$ s.t. $\text{VCD}(H^j, \Psi_{H^j}) \leq d - 1$.
(ii) $\forall j \in [m-1]$, $H^j$ is not empty and $H^j \{x_j\} = \{(1 - h_H(x_j))\}$.
(iii) $H^m = \{h_H\}$.

Lemma 4 suggests that for any $\mathcal{H}, \mathcal{X}$, one can partition the hypothesis space $\mathcal{H}$ into $m \leq |\mathcal{X}| + 1$ subspaces with lower VC dimension with respect to some compact-distinguishable set. The main idea of the lemma is similar to the reduction of a concept class w.r.t. some instance $x$ to lower VCD as done in Theorem 9 of [FW95]. The key distinction in our work is that we consider “compact-distinguishable” subsets of the instance space (as detailed in the proof sketch below), which in turn ensure the non-emptiness property stated in Lemma 4. Another key novelty in our proof of Theorem 3 is to recursively apply the reduction step.

To prove the lemma, we provide a constructive procedure to partition the hypothesis class, and show that the resulting partitions have reduced VC dimensions on some compact-distinguishable set. We highlight the procedure for constructing the partitions in Algorithm 2\cite{Algorithm 2}(Line 7–Line 10). In Figure 3 we provide an illustrative example for creating such partitions for the Warmuth hypothesis class as described in Table 2. We sketch the proof of Lemma 4 below, and defer the detailed proof to Appendix D.1 of the supplementary.

Proof [Proof Sketch of Lemma 4] Let us define $H_x = \{h \in H : h \triangle x \in \Psi_H \subset H \subset \Psi_H\}$. Here, $h \triangle x$ denotes the hypothesis that only differs with $h$ on the label of $x$. Fix a reference hypothesis $h_H$. $\forall x \in \Psi_H$, let $y_j = 1 - h_H(x_j)$ be the opposite label of $x_j$ as provided by $h_H$. As highlighted in Line 9 of Algorithm 2, we consider the set $H^1 := H^1_{x_1} = \{h \in H_{x_1} : h(x_1) = y_1\}$ as the first partition.

Here, we claim that $|H^1| > 0$. When $d = 1$, $\text{VCD}(H^1, \Psi_H \{x_1\}) = 0$ (see Appendix D.1 of the supplementary). When $d > 1$,

$$\text{VCD}(H^1, \Psi_H \{x_1\}) \leq \text{VCD}(H^1_{x_1}, \Psi_H) = \text{VCD}(H_{x_1}, \Psi_H) - 1 \leq \text{VCD}(H, \Psi_H) - 1 \leq d - 1$$

We further show that there exists a compact-distinguishable set $\Psi_H^{x_1} \subset \Psi_H \{x_1\}$ for the first partition $H^1$ (see Appendix D.1 of the supplementary). Then, we conclude that the first partition $H^1$ has $\text{VCD}(H^1, \Psi_{H^1}) \leq d - 1$.

Next, we remove the first partition $H^1$ from $H$, and continue to create the above mentioned partitions on $H_{\text{rest}} = H \setminus H^1$ and $X_{\text{rest}} = \Psi_H \{x_1\}$. As discussed in Appendix D.1 of the supplementary, we show that $X_{\text{rest}}$ is a compact-distinguishable set on $H_{\text{rest}}$. Therefore, we can repeat the above procedure (Line 7–Line 10, Algorithm 2) to create the subsequent partitions. This process continues until the size of $X_{\text{rest}}$ reduces to 1, i.e., $X_{\text{rest}} = \{x_{m-1}\}$. Until then, we obtain partitions $\{H^1, \ldots, H^{m-2}\}$. By construction, $H^j$ satisfy properties (i) and (ii) for all $j \in [m - 2]$.

It remains to show that the last two partitions $H^{m-1}$ and $H^m$ also satisfy properties (i) and (ii). Since $X_{\text{rest}} = \{x_{m-1}\}$ before we start iteration $m - 1$, there must exist exactly two hypotheses in $H_{\text{rest}}$, and therefore $|H^{m-1}|, |H^m| = 1$. This implies that $\text{VCD}(H^{m-1}, \Psi_{H^{m-1}}) = \text{VCD}(H^m, \Psi_{H^m}) = 0 \leq d - 1$. Furthermore, notice that for every $j \in [m - 1]$ and $h \in H^j, h_H(x_j) \neq h(x_j)$. This indicates $h_H \in H_m$. Since $|H_m| = 1$, we get $H_m = \{h_H\}$ which completes the proof.\hfill\qed
To begin with, we initialize

\[
\sigma_{\text{ws}}(h'; H, h) \leftarrow \begin{cases} 
0 & \text{if } h' = h \\
|h| + 1 & \text{o.w.} 
\end{cases}
\]

We then set the values of the preference function \(\text{VC dimension}\). This leads to the following proof for Theorem 3.

\[ (\text{Line 12}). \text{ Upon receiving } p \in V, \text{ the learner chooses } h' \text{ such that } \sigma_{\text{ws}}(h'; H, h) = V. \]

This indicates that none of the \(h_j \in H\) satisfies the three properties in Lemma 4. We show that for any target hypothesis \(h^*\), every teaching example \(x\) is distinguishable by \(\sigma_{\text{ws}}\), which induces uniform low preference, and there exists a teaching sequence of length up to \(d\) for any target hypothesis \(h^*\). We summarize the recursive procedure in Algorithm 2.

**Algorithm 2**: Recursive procedure for constructing \(\sigma_{\text{ws}}\) achieving \(T_{\text{ws}}(H, h_0, \sigma_{\text{ws}}) \leq \text{VCD}(H, X)\)

**Input**: \(H, \mathcal{X}, h_0\)

1. Let \(I : \mathcal{H} \rightarrow \{1, \ldots, |H|\}\) be any bijective mapping
2. \(\forall h, h' \in \mathcal{H}, H \subseteq \mathcal{H}: \)
   \[
   \sigma_{\text{ws}}(h'; H, h) \leftarrow \begin{cases} 
0 & \text{if } h' = h \\
|h| + 1 & \text{o.w.} 
\end{cases}
\]

3. **SetPreference** \((H, \mathcal{H}, \mathcal{X}, h_0)\)
4. **SetPreference** \((V, \hat{H}, \hat{X}, h)\)
5. Create compact-distinguishable set \(\hat{H} \subseteq \hat{X}\)
6. \(H \leftarrow \hat{H}, X \leftarrow \hat{X}\)
7. for \(x \in \hat{H}\) do
   8. \(H^* \leftarrow \{(h' \in H : h' \neq x, h'(x) = y)\}\)
   9. \(H \leftarrow H \cup H^*, X \leftarrow X \setminus \{x\}\)
10. \(V_{\text{next}} \leftarrow V \cap \hat{H}(\{(x, y)\})\)
11. for \(h' \in H^*\) do
   12. \(\sigma_{\text{ws}}(h'; V_{\text{next}}, h) \leftarrow I(h')\)
   13. \(h_{\text{next}} \leftarrow \arg \min_{h' \in H^*} I(h')\)
   14. **SetPreference** \((V_{\text{next}}, I_{\text{next}}, \hat{H}, \{x\}, h_{\text{next}})\)

**Recursive Construction of** \(\sigma_{\text{ws}}\) **achieving** \(T_{\text{ws}}(H, h_0, \sigma_{\text{ws}}) = \hat{O}(\text{VCD}(H, X))\)

In the following, we show that for any given \(X, \mathcal{H}, h_0\), one can recursively construct a preference function using the teaching examples, where the resulting teaching complexity is at most linear in the VC dimension. This leads to the following proof for Theorem 3.

**Proof**: [Proof of Theorem 3] In a nutshell, the proof consists of three steps: (i) Initialization of \(\sigma_{\text{ws}}\), (ii) setting the preferences by recursively invoking the constructive procedure for \(\text{Lemma 4}\), and (iii) showing that there exists a teaching sequence of length up to \(d\) for any target hypothesis \(h^*\). We summarize the recursive procedure in Algorithm 2.

**Step (i)**. To begin with, we initialize \(\sigma_{\text{ws}}\) with default values which induce uniform low preference, except for \(\sigma(h^*; H, h) = 0\) where \(h^* = h\) (cf. Line 2 of Algorithm 2). The self-preference guarantees that \(\sigma_{\text{ws}}\) is collusion-free as per Definition 1.

**Step (ii)**. The recursion begins at the top level with the set of “search space” \(H = \mathcal{H}\), current version space \(V = \mathcal{H}\), and initial hypothesis \(h = h_0\). **Lemma 4** suggests that we can divide \(H\) with examples in \(\hat{H}\) into \(m = |\hat{H}| + 1\) groups \(\{H^1, \ldots, H^m\}\) with the help of \(h\), where for all \(j \in [m]\), there exists a compact-distinguishable set \(\hat{H}_j\) that satisfies the three properties in **Lemma 4**.

Now consider some fixed hypothesis \(h \in H\). We show that for \(j \in [m - 1]\), every teaching example \((x_j, y_j)\), where \(x_j \in \hat{H}_j\) and \(y_j = 1 - h(x_j)\), corresponds to a unique version space \(V^j := \{h \in V : h(x_j) = y_j\}\). To prove this statement, we consider \(R^j := V^j \cap H = \{h \in H : h(x_j) = y_j\}\). As is discussed in Appendix D.2 of the supplementary, we know that none of \(R^j\) for \(j \in [m - 1]\) are equal. This indicates that none of \(V^j\) for \(j \in [m - 1]\) are equal either.

We then set the values of the preference function \(\sigma_{\text{ws}}(\cdot; V^j, h)\) for all \(j \in [m - 1]\) and \(y_j = 1 - h(x_j)\) (Line 12). Upon receiving \((x_j, y_j)\), the learner will be steered to the next “search space” \(H^j\), with version space \(V^j\). By **Lemma 4** we have \(\text{VCD}(H^j, \hat{H}_j) \leq \text{VCD}(H, \hat{H}) - 1\).
We will build the preference function \( \sigma_{\text{lyg}} \) recursively \( m - 1 \) times for each of the partitions \((H^j; \Psi_{H^j}; V^j)\) corresponding to \( j \in [m - 1] \) and \( y \in \{0, 1\} \). At each level of recursion, the VC dimension reduces by 1. We stop the recursion when \( \text{VCD}(H^j; \Psi_{H^j}) = 0 \). As Lemma 4 suggests, this corresponds to the scenario where there is only one hypothesis left.

**Step (iii).** Given the preference function constructed in Algorithm 2, we can build up the set of (labeled) teaching examples recursively. Let us consider the case where \( h^* \) belongs to the partition (of search spaces) corresponding to \( j = j^* \). Then, teacher provides an example given by \((x = x_{j^*}, y = h^*(x_{j^*}))\). After receiving the teaching example, the resulting partition \( H^j \) will stay in the version space; meanwhile, \( h_0 \) will be removed from version space. The new version space will be \( V^j \), and the new search space will be given by \( H^j \). Learner’s new hypothesis induced by the preference function is given by \( h = h_{\text{next}} \). By repeating this teaching process for a maximum of \( d \) examples, we reach the level where VC-dimension of the search space is 0. At this step \( h^* \) must be the only hypothesis left in the search space. So \( h_{\text{next}} = h^* \), and the learner has reached \( h^* \).

## 6 Discussion and Conclusion

We now discuss a few thoughts related to different families of preference functions. First of all, the size of the families grows exponentially as we change our model from \( \Sigma_{\text{const}} \), \( \Sigma_{\text{global}} \) to \( \Sigma_{\text{lyg}}/\Sigma_{\text{local}} \) and finally to \( \Sigma_{\text{lyg}} \), thus resulting in more powerful models with lower teaching complexity. While run time has not been the focus of this paper, it is an interesting direction of future work to understand the run time complexity of teaching models under complex preferences. For instance, as the size of the families grow, the problem of finding the best preference function \( \sigma \) in a given family \( \Sigma \) that achieve the minima in Eq. (3.2) becomes more computationally challenging. Furthermore, it is an interesting to characterize the presumably increased run time complexity of sequential learners and teachers with complex preference functions.

The recursive procedure in Algorithm 2 creates a preference function \( \sigma_{\text{lyg}} \in \Sigma_{\text{lyg}} \) that has teaching complexity at most \( \text{VCD} \). It is interesting to note that the resulting preference function \( \sigma_{\text{lyg}} \) has the characteristic of "win-stay, loose shift" \([\text{BDGG14, CSMA18}]\). Given that for any hypothesis we have \( \sigma(h^*, h) = 0 \), the learner prefers her current hypothesis as long as it remains consistent. Preference functions with this characteristic naturally exhibit the "collusion-free" property (Definition 1). For some problems, one can achieve even lower teaching complexity compared to \( \text{VCD} \) for the \( \Sigma_{\text{lyg}} \) family which do not have the "win-stay, loose shift" characteristic but still satisfy the "collusion-free" property (Definition 1). In fact, the teaching sequence we provided for the Warmuth class in Table 2a is one such example.

One fundamental aspect of modeling teacher-learner interactions is the notion of collusion-free teaching. Collusion-freeness for the batched setting is well established in the research community and \( \text{NCTD} \) characterizes the complexity of the strongest collusion-free batch model. The implication of collusion-freeness is more subtle and less understood under the sequential setting. In this paper, we are introducing a possible notion of collusion-freeness for the sequential setting (Definition 1). As discussed above, a stricter condition is the "win-stay lose-shift" model, which is easier to validate without running the teaching algorithm. In contrast, the condition of Definition 1 is more involved in terms of validation and is a joint property of the teacher-learner pair. One intriguing question for future work is defining notions of collusion-free teaching in sequential models and understanding their implications on teaching complexity.

Another interesting direction of future work is to better understand the properties of the teaching parameter \( \Sigma_{\text{TD}} \). One question of particular interest is showing that the teaching parameter is not upper bounded by any constant independent of the hypothesis class, which would suggest that a strong collusion in our model. We can show that for certain hypothesis classes, \( \Sigma_{\text{TD}} \) is lower bounded by a function of \( \text{VCD} \). In particular, for the power set class of size \( d \) (which has \( \text{VCD} = d \)), \( \Sigma_{\text{TD}} \) is lower bounded by \( \Omega \left( \frac{d}{\log d} \right) \). Another direction of future work is to understand whether this parameter is additive or subadditive over disjoint domains. Also, we consider a generalization to the infinite VC classes as a very interesting direction for future work.

Our framework provides novel tools for reasoning about teaching complexity by constructing preference functions. This opens up an interesting direction of research to tackle important open problems, such as proving whether \( \text{NCTD} \) or \( \text{RTD} \) is linear in \( \text{VCD} \) \([\text{SZ15, CCT16, HWL17, KSZ19}]\). In this paper, we showed that neither of the families \( \Sigma_{\text{lyg}} \) and \( \Sigma_{\text{local}} \) dominates the other (Theorem 2). As a direction for future work, it would be important to further quantify the complexity of \( \Sigma_{\text{local}} \) families.
Acknowledgements

Xiaojin Zhu is supported by NSF 1545481, 1561512, 1623605, 1704117, 1836978 and the MADLab AF CoE FA9550-18-1-0166.

References


