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 $\sigma \in \Sigma$ induces a teacher-learner pair with teaching complexity as $\mathsf{TD}(\sigma)$. 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, KDCS19], models inspired by control theory [Zhu18, LZZ19], models

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

for sequential tasks [CL12, HTS18, TGH$^+$19], and models for human-centered applications that require adaptivity [SBB$^+$13, HCMA$^+$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 Σ. 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 σ induces a teacher-learner pair with teaching dimension $\mathsf{TD}(\sigma)$ (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 σ functions in our framework (see §4 and Table 1).
- We study the differences in the family of σ functions inducing the strongest batch model [KSZ19] and functions inducing a weak sequential model [CSMA$^+$18] (§5.2) (also, see the relationship between Σ_{gvs} and Σ_{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 σ 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.

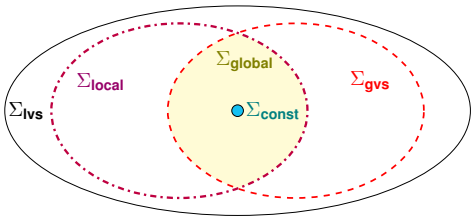


Figure 1: Venn diagram for different families of preference functions.

Families	Σ_{const}	Σ_{global}	Σ_{gvs}	Σ_{local}	Σ_{lvs}
Reduction	TD	RTD / PBTD	NCTD	Local-PBTD	–
Complexity Results	–	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD})$
	[GK95]	[ZLHZ11, GRSZ17, HWLW17]	[KSZ19]	[CSMA$^+$18]	

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 $h : \mathcal{X} \to \mathcal{Y}$. 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 $h^\star \in \mathcal{H}$ 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 $h^\star$, i.e., $y_z = h^\star(x_z)$. 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., $\mathcal{H}(Z) := \{h \in \mathcal{H} \mid \forall z = (x_z, y_z) \in Z, h(x_z) = y_z\}$.

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 $\sigma : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R}$. For any two hypotheses h', h'', we say that the learner prefers h' to h'' based on the current hypothesis h and version space $H \subseteq \mathcal{H}$, iff $\sigma(h'; H, h) < \sigma(h''; H, h)$. If $\sigma(h'; H, h) = \sigma(h''; H, h)$, then the learner could pick either one of these two.

Interaction protocol and teaching objective. The teacher's goal is to steer the learner towards the target hypothesis $h^\star$ 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 $z_t \in \mathcal{Z}$, 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 σ:

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

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 equals $h^\star$. We summarize the interaction in Protocol 1.[1]

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** $h_t = h^\star$ **then** teaching process terminates

3 The Complexity of Teaching

3.1 Teaching Dimension for a Fixed Preference Function

Our objective is to design teaching algorithms that can steer the learner towards the target hypothesis in a minimal number of time steps. We study the *worst-case* number of steps needed, as is common when measuring information complexity of teaching [GK95, ZLHZ11, GRSZ17, Zhu18]. Fix the ground set of instances $\mathcal{X}$ and the learner's preference σ. For any version space $H \subseteq \mathcal{H}$, the worst-case optimal cost for steering the learner from h to $h^\star$ is characterized by

$$D_\sigma(H, h, h^\star) = \begin{cases} 1, & \exists z, \text{ s.t. } \mathbf{C}_\sigma(H, h, z) = \{h^*\} \\ 1 + \min_z \max_{h'' \in \mathbf{C}_\sigma(H,h,z)} D_\sigma(H \cap \mathcal{H}(\{z\}), h'', h^\star), & \text{otherwise} \end{cases}$$

where $\mathbf{C}_\sigma(H, h, z) = \arg\min_{h' \in H \cap \mathcal{H}(\{z\})} \sigma(h'; H \cap \mathcal{H}(\{z\}), h)$ denotes the set of candidate hypotheses most preferred by the learner. Note that our definition of teaching dimension is similar in spirit to the local preference-based teaching complexity defined by [CSMA$^+$18]. We shall see in the next section, this complexity measure in fact reduces to existing notions of teaching complexity for specific families of preference functions.

Given a preference function σ and the learner's initial hypothesis h_0, the teaching dimension w.r.t. σ is defined as the worst-case optimal cost for teaching any target $h^\star$:

$$\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma) = \max_{h^\star} D_\sigma(\mathcal{H}, h_0, h^\star). \tag{3.1}$$

3.2 Teaching Dimension for a Family of Preference Functions

In this paper, we will investigate several families of preference functions (as illustrated in Figure 1). For a family of preference functions Σ, we define the teaching dimension w.r.t the family Σ as the teaching dimension w.r.t. the *best* σ in that family:

$$\Sigma\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \min_{\sigma \in \Sigma} \mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma). \tag{3.2}$$

[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]).

3.3 Collusion-free 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, the notion of collusion-free teaching in the sequential models is not well understood. We introduce a novel notion of collusion-freeness for the sequential setting, which captures the following idea: if h is the only hypothesis in the most preferred set defined by σ, then the learner will always stay at h as long as additional information received by the learner is consistent with h. We formalize this notion in the definition below. Note that for σ functions corresponding to batch models (see §4), Definition 1 reduces to the collusion-free definition of [GM96].

Definition 1 (Collusion-free preference) *Consider a time t where the learner's current hypothesis is h_{t-1} and version space is H_t (see Protocol 1). Further assume that the learner's preferred hypothesis for time t is uniquely given by $\arg\min_{h' \in H_t} \sigma(h'; H_t, h_{t-1}) = \{\hat{h}\}$. Let S be additional examples provided by an adversary from time t onwards. We call a preference function collusion-free, if for any S consistent with $\hat{h}$, it holds that $\arg\min_{h' \in H_t \cap \mathcal{H}(S)} \sigma(h'; H_t \cap \mathcal{H}(S), \hat{h}) = \{\hat{h}\}$.*

In this paper, we study preference functions that are collusion-free. In particular, we use Σ_{CF} to denote the set of preference functions that induce collusion-free teaching:

$$\Sigma_{\mathsf{CF}} = \{\sigma \mid \sigma \text{ is collusion-free}\}.$$

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 Σ_{const}, which corresponds to constant preference functions:

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

The second family, denoted by Σ_{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:

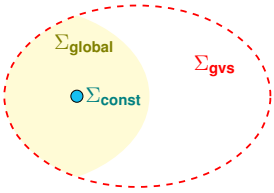


Figure 2: Batch models.

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

The third family, denoted by Σ_{gvs}, corresponds to the preference functions that depend on the learner's version space, but do not depend on the learner's current hypothesis:

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

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 existing notions of teaching dimension.

Definition 2 (Vapnik–Chervonenkis dimension [VC71]) *The VC dimension for $H \subseteq \mathcal{H}$ w.r.t. a fixed set of unlabeled instances $X \subseteq \mathcal{X}$, denoted by $\mathsf{VCD}(H, X)$, is the cardinality of the largest set of points $X' \subseteq X$ that are "shattered".[2] Formally, let $H_{|X} = \{(h(x_1), ..., h(x_n)) \mid \forall h \in H\}$ denote all possible patterns of H on X. Then $\mathsf{VCD}(H, X) = \max |X'|$, s.t. $X' \subseteq X$ and $|H_{|X'}| = 2^{|X'|}$.*

[2] In the classical definition of VCD, only the first argument H is present; the second argument X is omitted and is by default the ground set of unlabeled instances $\mathcal{X}$.

Definition 3 (Teaching dimension [GK95]) *For any hypothesis $h \in \mathcal{H}$, we call a set of instances $\mathrm{T}(h) \subseteq \mathcal{X}$ a teaching set for h, if it can uniquely identify $h \in \mathcal{H}$. The teaching dimension for $\mathcal{H}$, denoted by $\mathsf{TD}(\mathcal{H})$, is the maximum size of the minimum teaching set for any $h \in \mathcal{H}$: $\mathsf{TD}(\mathcal{H}) = \max_{h\in\mathcal{H}} \min |\mathrm{T}(h)|$.*

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, ZLHZ11]) *The recursive teaching dimension (RTD) of $\mathcal{H}$, denoted by $\mathsf{RTD}(\mathcal{H})$, is the smallest number k, such that one can find an ordered sequence of hypotheses in $\mathcal{H}$, denoted by $(h_1, \ldots, h_i, \ldots, h_{|\mathcal{H}|})$, where every hypothesis h_i has a teaching set of size no more than k to be distinguished from the hypotheses in the remaining sequence.*

In this paper we consider finite hypothesis classes. Under this setting, RTD is equivalent to preference-based teaching dimension (PBTD) [GRSZ17].

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 $\mathcal{H}$ be a hypothesis class and $\mathrm{T} : \mathcal{H} \to 2^{\mathcal{X}}$ be a "teacher mapping" on $\mathcal{H}$, i.e., mapping a given hypothesis to a teaching set.*[3] *We say that T is non-clashing on $\mathcal{H}$ iff there are no two distinct $h, h' \in \mathcal{H}$ such that $\mathrm{T}(h)$ is consistent with h' and $\mathrm{T}(h')$ is consistent with h. The non-clashing Teaching Dimension of $\mathcal{H}$, denoted by $\mathsf{NCTD}(\mathcal{H})$, is defined as $\mathsf{NCTD}(\mathcal{H}) = \min_{\mathrm{T} \text{ is non-clashing}} \{\max_{h\in\mathcal{H}} |\mathrm{T}(h)|\}$.*

We show in the following, that the teaching dimension Σ-TD in Eq. (3.2) unifies the above definitions of TD's for batch models.

Theorem 1 (Reduction to existing notions of TD's) *Fix $\mathcal{X}, \mathcal{H}, h_0$. The teaching complexity for the three families reduces to the existing notions of teaching dimensions:*

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

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

In Table 2, we consider the well known Warmuth hypothesis class [DFSZ14] where $\Sigma_{\mathsf{const}}\text{-}\mathsf{TD} = 3$, $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD} = 3$, and $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD} = 2$. Table 2b and Table 2d show preference functions $\sigma \in \Sigma_{\mathsf{const}}$, $\sigma \in \Sigma_{\mathsf{global}}$, and $\sigma \in \Sigma_{\mathsf{gvs}}$ that achieve the minima in Eq. (3.2). Table 2a shows the teaching sequences achieving these teaching dimensions for these preference functions. In Appendix A.1, we provide another hypothesis class where $\Sigma_{\mathsf{const}}\text{-}\mathsf{TD} = 3$, $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD} = 2$, and $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD} = 1$.

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-1}. The first one is the family of local preference-based functions [CSMA+18], denoted by Σ_{local}, which corresponds to preference functions that depend on the learner's current (local) hypothesis, but do not depend on the learner's version space:

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

[3] We refer the reader to the original paper [KSZ19] for a more formal description of "teacher mapping".

h \ x	x_1	x_2	x_3	x_4	x_5	$\mathcal{S}_{\mathsf{const}} = \mathcal{S}_{\mathsf{global}}$	$\mathcal{S}_{\mathsf{gvs}}$	$\mathcal{S}_{\mathsf{local}}$	$\mathcal{S}_{\mathsf{lvs}}$
h_1	1	1	0	0	0	(x_1, x_2, x_4)	(x_1, x_2)	(x_1)	(x_1)
h_2	0	1	1	0	0	(x_2, x_3, x_5)	(x_2, x_3)	(x_3)	(x_2)
h_3	0	0	1	1	0	(x_1, x_3, x_4)	(x_3, x_4)	(x_3, x_4)	(x_3)
h_4	0	0	0	1	1	(x_2, x_4, x_5)	(x_4, x_5)	(x_5, x_4)	(x_4)
h_5	1	0	0	0	1	(x_1, x_3, x_5)	(x_1, x_5)	(x_5)	(x_5)
h_6	1	1	0	1	0	(x_1, x_2, x_4)	(x_2, x_4)	(x_4)	(x_3)
h_7	0	1	1	0	1	(x_2, x_3, x_5)	(x_3, x_5)	(x_3, x_5)	(x_4)
h_8	1	0	1	1	0	(x_1, x_3, x_4)	(x_1, x_4)	(x_4, x_3)	(x_5)
h_9	0	1	0	1	1	(x_2, x_4, x_5)	(x_2, x_5)	(x_4, x_5)	(x_1)
h_{10}	1	0	1	0	1	(x_1, x_3, x_5)	(x_1, x_3)	(x_5, x_3)	(x_2)

(a) The Warmuth hypothesis class and the corresponding teaching sequences (denoted by $\mathcal{S}$).

h'	$\forall h' \in H$
$\sigma_{\mathsf{const}}(h'; \cdot, \cdot)$ $\sigma_{\mathsf{global}}(h'; \cdot, \cdot)$	0

(b) σ_{const} and σ_{global}

$h \backslash h'$	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
$\sigma_{\mathsf{local}}(h'; \cdot, h = h_1)$	0	2	4	4	2	1	3	3	3	3
...										

(c) σ_{local} representing the Hamming distance between h' and h.

h'	h_1	h_2	...
H	$\{h_1, h_6\}$	$\{h_2, h_7\}$	...
	$\{h_1\}$	$\{h_2\}$	...
σ_{gvs}	0	0	...

(d) $\sigma_{\mathsf{gvs}}(h'; H, \cdot)$

h'	h_1	h_2	...
H	$\{h_1\} \cup$ $\{h_5, h_6, h_8, h_{10}\}^*$	$\{h_2\} \cup$ $\{h_1, h_7, h_6, h_9\}^*$	... ...
h	h_1	h_1 h_2	...
σ_{lvs}	0	0 0	...

(e) $\sigma_{\mathsf{lvs}}(h'; H, h)$. Here, $\{\cdot\}^*$ denotes all subsets.

Table 2: Teaching sequences with different preference functions for the Warmuth hypothesis class [DFSZ14].[4] Full preference functions are given in Appendix B.

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

$$\Sigma_{\mathsf{lvs}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists\, g : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R},\ \text{s.t.}\ \forall h', H, h, \sigma(h'; H, h) = g(h', H, h)\}$$

Figure 1 illustrates the relationship between these preference families. As an example, in Table 2c and Table 2e, we provide the preference functions σ_{local} and σ_{lvs} for the Warmuth hypothesis class that achieve the minima in Eq. (3.2).

5.2 Comparing Σ_{gvs}-TD and Σ_{local}-TD

In the following, we show that substantial differences arise as we transition from σ functions inducing the strongest batch (i.e., non-clashing) model to σ functions inducing a weak sequential (i.e., local preference-based) model. We provide the full proof of Theorem 2 in Appendix C.

Theorem 2 *Neither of the families Σ_{gvs} and Σ_{local} dominates the other. Specifically,*

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

5.3 Complexity Results

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

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

To establish the proof, we first introduce an important definition (Definition 6) and a key lemma (Lemma 4).

[4]The Warmuth hypothesis class is the smallest concept class for which RTD exceeds VCD.

Definition 6 (Compact-Distinguishable Set) *Fix $H \subseteq \mathcal{H}$ and $X \subseteq \mathcal{X}$, where $X = \{x_1, ..., x_n\}$. Let $H_{|X} = \{(h(x_1), ..., h(x_n)) \mid \forall h \in H\}$ 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 Ψ_H to denote a compact-distinguishable set on H.*

In words, one can uniquely identify any hypothesis in H with a (sub)set of examples from Ψ_H (also see the definition of distinguishing sets in [DFSZ14]). Our definition of compact-distinguishable set further implies that there are no "redundant" examples in Ψ_H. It can be shown that a compact-distinguishable set satisfies the following two properties: (i) it does not contain any pair of distinct instances 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 instance x such that $(\forall h \in H : h(x) = 1)$ or $(\forall h \in H : h(x) = 0)$.

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

(i) $\forall j \in [m]$*, there exists a compact-distinguishable set Ψ_{H^j} s.t. $\mathsf{VCD}(H^j, \Psi_{H^j}) \leqslant d - 1$.*

(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 class $\mathcal{H}$ into $m \leqslant |\mathcal{X}| + 1$ subsets with lower VC dimension with respect to some compact-distinguishable set.[5] 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 of Lemma 4 is that we consider compact-distinguishable sets for this partitioning, which in turn ensures the uniqueness of the version spaces associated with these partitions (see proof of Theorem 3). Another key novelty in our proof of Theorem 3 is to recursively apply the reduction step from the lemma.

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 (Line 7– Line 10). In Figure 3, we provide an illustrative example for creating such partitions for the Warmuth hypothesis class from Table 2a. We sketch the proof of Lemma 4 below, and defer the detailed proof to Appendix D.1.

Proof [Proof Sketch of Lemma 4] Let us define $H_x = \{h \in H : h \triangle x_{|\Psi_H} \in H_{|\Psi_H}\}$. Here, $h \triangle x$ denotes the hypothesis that only differs with h on the label of x, and $h_{|\Psi_H}$ denotes the patterns of h on Ψ_H. Fix a reference hypothesis h_H. For all $j \in [m-1]$, let $y_j = 1 - h_H(x_j)$ be the opposite label of $x_j \in \Psi_H$ as provided by h_H. As shown in Line 9 of Algorithm 2, we consider the set $H^1 := H^{y_1}_{x_1} = \{h \in H_{x_1} : h(x_1) = y_1\}$ as the first partition. In the appendix, we show that $|H^1| > 0$.

Next, we show that $\mathsf{VCD}(H^1, \Psi_H \backslash \{x_1\}) \leqslant d - 1$. When $d > 1$, we prove the statement as follows:

$$\mathsf{VCD}(H^1, \Psi_H \backslash \{x_1\}) \leqslant \mathsf{VCD}(H^{y_1}_{x_1}, \Psi_H) = \mathsf{VCD}(H_{x_1}, \Psi_H) - 1 \leqslant \mathsf{VCD}(H, \Psi_H) - 1 \leqslant d - 1$$

In the appendix, we prove the statement for $d = 1$, and further show that there exists a compact-distinguishable set $\Psi_{H^1} \subseteq \Psi_H \backslash \{x_1\}$ for the first partition H^1. Then, we conclude that the first partition H^1 has $\mathsf{VCD}(H^1, \Psi_{H^1}) \leqslant d - 1$.

Next, we remove the first partition H^1 from H, and continue to create the above mentioned partitions on $H_{\text{rest}} = H \backslash H^1$ and $X_{\text{rest}} = \Psi_H \backslash \{x_1\}$. As discussed in the appendix, we show that X_{rest} is a compact-distinguishable set on H_{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_{rest} reduces to 1, i.e. $X_{\text{rest}} = \{x_{m-1}\}$. Until then, we obtain partitions $\{H^1, ..., H^{m-2}\}$. By construction, H^j satisfy properties (i) and (ii) for all $j \in [m-2]$.

It remains to show that H^{m-1} and H^m also satisfy the properties in Lemma 4. Since $X_{\text{rest}} = \{x_{m-1}\}$ before we start iteration $m-1$, and X_{rest} is a compact-distinguishable set for H_{rest}, there must exist exactly two hypotheses in H_{rest}, and therefore $|H^{m-1}|, |H^m| = 1$. This implies that $\mathsf{VCD}(H^{m-1}, \Psi_{H^{m-1}}) = \mathsf{VCD}(H^m, \Psi_{H^m}) = 0$. Furthermore, $\forall j \in [m-1]$ and $h \in H^j$, we have $h_H(x_j) \neq h(x_j)$. This indicates $h_H \in H_m$, and hence $H_m = \{h_H\}$ which completes the proof. ■

[5] When $\mathsf{VCD}(H, \Psi_H) = 0$, this implies $|H| = 1$.

Algorithm 2 Recursive procedure for constructing σ_{lvs} achieving $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma_{\mathsf{lvs}}) \leqslant \mathsf{VCD}(\mathcal{H},\mathcal{X})$

Input: $\mathcal{X}$, $\mathcal{H}$, h_0

1: Let $I : \mathcal{H} \to \{1, \ldots, |\mathcal{H}|\}$ be any bijective mapping
2: For all $h' \in \mathcal{H}$, $H \subseteq \mathcal{H}$, $h \in \mathcal{H}$, initialize

$$\sigma_{\mathsf{lvs}}(h'; H, h) \leftarrow \begin{cases} 0 & \text{if } h' = h \\ |\mathcal{H}| + 1 & \text{o.w.} \end{cases}$$

3: SETPREFERENCE$(\mathcal{H}, \mathcal{H}, \mathcal{X}, h_0)$
4: **function** SETPREFERENCE(V, H, X, h)
5: Create compact-distinguishable set $\Psi_H \subseteq X$
6: $H_{\text{rest}} := H, X_{\text{rest}} := \Psi_H$
7:
8:
9:
10:
11: $V_{\text{next}} \leftarrow V \cap \mathcal{H}(\{(x, y)\})$
12: **for** $h' \in H_x^y$ **do** $\sigma_{\mathsf{lvs}}(h'; V_{\text{next}}, h) \leftarrow I(h')$
13: $h_{\text{next}} \leftarrow \arg\min_{h' \in H_x^y} I(h')$
14: SETPREFERENCE$(V_{\text{next}}, H_x^y, \Psi_H \backslash \{x\}, h_{\text{next}})$

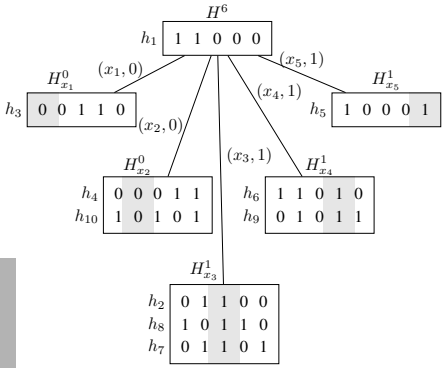


Figure 3: Illustration of Lemma 4 on the Warmuth class. The grouped hypotheses in the leaf clusters correspond to the sets H_x^y created in Line 9 of Algorithm 2.

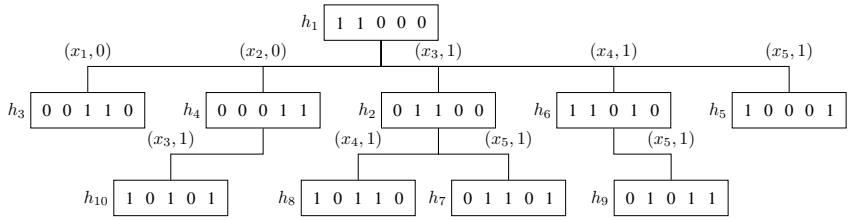


Figure 4: Illustration of Theorem 3 proof – constructing a $\sigma_{\mathsf{lvs}} \in \Sigma_{\mathsf{lvs}}$ for the Warmuth class.

Recursive construction of σ_{lvs}. As a part of the Theorem 3 proof, we provide a recursive procedure for constructing a $\sigma_{\mathsf{lvs}} \in \Sigma_{\mathsf{lvs}}$ achieving $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma_{\mathsf{lvs}}) = O\left(\mathsf{VCD}(\mathcal{H},\mathcal{X})\right)$.

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

Step (i). To begin with, we initialize σ_{lvs} with default values which induce high σ values (i.e., low preference), except for $\sigma(h'; H, h) = 0$ where $h' = h$ (c.f. Line 2 of Algorithm 2). The self-preference guarantees that σ_{lvs} is collusion-free as per Definition 1.

Step (ii). The recursion begins at the top level with $H = \mathcal{H}$, current version space $V = \mathcal{H}$, and initial hypothesis $h = h_0$. Lemma 4 suggests that we can partition H into $m = |\Psi_H| + 1$ groups $\{H^1, ..., H^m\}$, where for all $j \in [m]$, there exists a compact-distinguishable set Ψ_{H^j} that satisfies the properties in Lemma 4.

Now consider the hypothesis $h := h_0$. We show that for $j \in [m-1]$, every (x_j, y_j), where $x_j \in \Psi_H$ 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 $R^j := V^j \cap H = \{h \in H : h(x_j) = y_j\}$. According to Lemma 8 of Appendix D.2, 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.

We then set the values of the preference function $\sigma_{\mathsf{lvs}}(\cdot; V^j, h)$ 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 $\mathsf{VCD}(H^j, \Psi_{H^j}) \leqslant \mathsf{VCD}(H, \Psi_H) - 1$.

We will build the preference function σ_{lvs} recursively $m - 1$ times for each $(V^j, H^j, \Psi_{H^j}, h_{\text{next}})$, where h_{next} corresponds to the unique hypothesis identified by function I (Line 13–Line 14). At

each level of recursion, VCD reduces by 1. We stop the recursion when $\mathsf{VCD}(H^j; \Psi_{H^j}) = 0$, which corresponds to the scenario $|H^j| = 1$.

Step (iii). Given the preference function constructed in Algorithm 2, we can build up the set of (labeled) teaching examples recursively. Consider the beginning of the teaching process, where the learner's current hypothesis is h_0 and version space is $\mathcal{H}$, and the goal of the teacher is to teach $h^\star$. Consider the first level of the recursion in Algorithm 2, where we divide $\mathcal{H}$ into $m = |\Psi_{\mathcal{H}}| + 1$ groups $\{H^1, ..., H^m\}$. Let us consider the case where $h^\star \in H^{j^\star}$ with $j^\star \in [m-1]$. The teacher provides an example given by $(x = x_{j^\star}, y = h^\star(x_{j^\star}))$. After receiving the teaching example, the resulting partition $H^{j^\star}$ will stay in the version space; meanwhile, h_0 will be removed from the version space. The new version space will be $V^{j^\star}$. The learner's new hypothesis induced by the preference function is given by $h_{\text{next}} \in H^{j^\star}$. By repeating this teaching process for a maximum of d steps, the learner reaches a partition of size 1 (see *Step (ii)* for details). At this step $h^\star$ must be the only hypothesis left in the search space. Therefore, $h_{\text{next}} = h^\star$, and the learner has reached $h^\star$. ■

Figure 4 illustrates the recursive construction of a $\sigma_{\mathsf{lvs}} \in \Sigma_{\mathsf{lvs}}$ for the Warmuth class, with $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma_{\mathsf{lvs}}) = 2$.

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 Σ_{const}, Σ_{global} to Σ_{gvs}/Σ_{local} and finally to Σ_{lvs}, thus resulting in more powerful models with lower teaching complexity. While run time has not been the focus of this paper, it would be interesting to characterize the presumably increased run time complexity of sequential learners and teachers with complex preference functions. Furthermore, as the size of the families grow, the problem of finding the best preference function σ in a given family Σ that achieve the minima in Eq. (3.2) becomes more computationally challenging.

The recursive procedure in Algorithm 2 creates a preference function $\sigma_{\mathsf{lvs}} \in \Sigma_{\mathsf{lvs}}$ that has teaching complexity at most VCD. It is interesting to note that the resulting preference function σ_{lvs} has the characteristic of "win-stay, loose shift" [BDGG14, CSMA+18]: Given that for any hypothesis we have $\sigma(h; \cdot, h) = 0$, the learner prefers her current hypothesis as long as it remains consistent. Preference functions with this characteristic naturally exhibit the collusion-free property in Definition 1. For some problems, one can achieve lower teaching complexity for a $\sigma \in \Sigma_{\mathsf{lvs}}$. In fact, the preference function σ_{lvs} we provided for the Warmuth class in Table 2e has teaching complexity 1, while the preference function constructed in Figure 4 has teaching complexity 2.

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 NCTD characterizes the complexity of the strongest collusion-free batch model. In this paper, we are introducing a new 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 Σ-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 a strong collusion in our model. We can show that for certain hypothesis classes, Σ-TD is lower bounded by a function of VCD. In particular, for the power set class of size d (which has $\mathsf{VCD} = d$), Σ-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 of our results 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 NCTD or RTD is linear in VCD [SZ15, CCT16, HWLW17, KSZ19]. In this paper, we showed that neither of the families Σ_{gvs} and Σ_{local} dominates the other (Theorem 2). As a direction for future work, it would be important to further quantify the complexity of Σ_{local} family.

Acknowledgements

This work was done in part when Yuxin Chen was at Caltech. Xiaojin Zhu is supported by NSF 1545481, 1561512, 1623605, 1704117, 1836978 and the MADLab AF CoE FA9550-18-1-0166.

References

[AK97] Dana Angluin and Mārtiņš Kriķis. Teachers, learners and black boxes. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 285–297. ACM, 1997.

[BDGG14] Elizabeth Bonawitz, Stephanie Denison, Alison Gopnik, and Thomas L Griffiths. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, 74:35–65, 2014.

[CCT16] Xi Chen, Yu Cheng, and Bo Tang. On the recursive teaching dimension of vc classes. In *Advances in Neural Information Processing Systems*, pages 2164–2171, 2016.

[CL12] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI*, 2012.

[CSMA+18] Yuxin Chen, Adish Singla, Oisin Mac Aodha, Pietro Perona, and Yisong Yue. Understanding the role of adaptivity in machine teaching: The case of version space learners. In *Advances in Neural Information Processing Systems*, pages 1476–1486, 2018.

[DFSZ14] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, vc-dimension and sample compression. *JMLR*, 15(1):3107–3131, 2014.

[FW95] Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[GK95] Sally A Goldman and Michael J Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

[GM96] Sally A Goldman and H David Mathias. Teaching a smarter learner. *Journal of Computer and System Sciences*, 52(2):255–267, 1996.

[GRSZ17] Ziyuan Gao, Christoph Ries, Hans U Simon, and Sandra Zilles. Preference-based teaching. *JMLR*, 18(31):1–32, 2017.

[HCMA+19] Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. Teaching multiple concepts to a forgetful learner. In *Advances in Neural Information Processing Systems*, 2019.

[HTS18] Luis Haug, Sebastian Tschiatschek, and Adish Singla. Teaching inverse reinforcement learners via features and demonstrations. In *Advances in Neural Information Processing Systems*, pages 8464–8473, 2018.

[HWLW17] Lunjia Hu, Ruihan Wu, Tianhong Li, and Liwei Wang. Quadratic upper bound for recursive teaching dimension of finite VC classes. In *Proceedings of the 30th Conference on Learning Theory, COLT*, pages 1147–1156, 2017.

[KDCS19] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive teaching algorithms for inverse reinforcement learning. In *IJCAI*, pages 2692–2700, 2019.

[KKW07] Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 09 2007.

[KSZ19] David Kirkpatrick, Hans U. Simon, and Sandra Zilles. Optimal collusion-free teaching. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 506–528, 2019.

[Kuh99] Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *European Conference on Computational Learning Theory*, pages 168–182. Springer, 1999.

[LDH+17] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. Iterative machine teaching. In *ICML*, pages 2149–2158, 2017.

[LDL+18] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M. Rehg, and Le Song. Towards black-box iterative machine teaching. In *ICML*, pages 3147–3155, 2018.

[LZZ19] Laurent Lessard, Xuezhou Zhang, and Xiaojin Zhu. An optimal control approach to sequential machine teaching. In *AISTATS*, pages 2495–2503, 2019.

[OS99] Matthias Ott and Frank Stephan. Avoiding coding tricks by hyperrobust learning. In *European Conference on Computational Learning Theory*, pages 183–197. Springer, 1999.

[SBB+13] Adish Singla, Ilija Bogunovic, G Bartók, A Karbasi, and A Krause. On actively teaching the crowd to classify. In *NIPS Workshop on Data Driven Education*, 2013.

[SBB+14] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, pages 154–162, 2014.

[SZ15] Hans U Simon and Sandra Zilles. Open problem: Recursive teaching dimension versus vc dimension. In *Conference on Learning Theory*, pages 1770–1772, 2015.

[TGH+19] Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems*, 2019.

[VC71] VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

[Zhu13] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems*, pages 1905–1913, 2013.

[Zhu15] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pages 4083–4087, 2015.

[Zhu18] Xiaojin Zhu. An optimal control view of adversarial machine learning. *arXiv preprint arXiv:1811.04422*, 2018.

[ZLHZ08] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In *COLT*, pages 135–146, 2008.

[ZLHZ11] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *JMLR*, 12(Feb):349–384, 2011.

[ZSZR18] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An overview of machine teaching. *CoRR*, abs/1801.05927, 2018.

A Supplementary Materials for §4

A.1 An Example Hypothesis Class and the Teaching Sequences for the Batch Models

In this section, we provide an example hypothesis class where $\Sigma_{\mathsf{const}}\text{-}\mathsf{TD} = \mathsf{TD} = 3$, $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD} = \mathsf{RTD} = 2$, and $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD} = \mathsf{NCTD} = 1$. The hypothesis class is specified in Table 3a. The preference functions inducing the optimal teaching sets for the examples are specified in Table 3b, 3c, and 3d.

$\mathcal{H}$ \ $\mathcal{X}$	x_1	x_2	x_3	x_4	x_5	x_6	$\mathcal{S}_{\mathsf{const}}$	$\mathcal{S}_{\mathsf{global}}$	$\mathcal{S}_{\mathsf{gvs}}$
h_1	1	0	0	0	0	1	(x_1, x_6)	(x_1, x_6)	(x_1)
h_2	0	1	0	0	0	1	(x_2, x_6)	(x_2, x_6)	(x_2)
h_3	1	1	1	0	0	0	(x_3, x_4, x_5)	(x_1)	(x_3)
h_4	1	1	1	1	0	0	(x_4, x_5)	(x_4, x_5)	(x_4)
h_5	1	1	1	0	1	0	(x_4, x_5)	(x_4, x_5)	(x_5)
h_6	0	0	0	1	1	1	(x_4, x_5)	(x_4, x_5)	(x_6)

(a) An example hypothesis class with the optimal teaching sets under different families of preference functions.

h'	h_1	h_2	h_3	h_4	h_5	h_6
$\sigma_{\mathsf{const}}(h'; \cdot, \cdot)$	0	0	0	0	0	0

(b) Preference function σ_{const}

h'	h_1	h_2	h_3	h_4	h_5	h_6
$\sigma_{\mathsf{global}}(h'; \cdot, \cdot)$	1	1	0	1	1	1

(c) Preference function σ_{global}

h'	h_1	h_2	h_3	h_4	h_5	h_6
H	$\{h_1, h_3, h_4, h_5\}$ $\{h_1, h_3, h_4\}$ $\{h_1, h_3, h_5\}$ $\{h_1\}$	$\{h_2, h_3, h_4, h_5\}$ $\{h_2, h_3, h_4\}$ $\{h_2, h_3, h_5\}$ $\{h_2\}$	$\{h_3, h_4, h_5\}$ $\{h_3, h_4\}$ $\{h_3, h_5\}$ $\{h_3\}$	$\{h_4, h_6\}$ $\{h_4\}$	$\{h_5, h_6\}$ $\{h_5\}$	$\{h_1, h_2, h_6\}$ $\{h_1, h_6\}$ $\{h_2, h_6\}$ $\{h_6\}$
$\sigma_{\mathsf{gvs}}(h'; H, \cdot)$	0	0	0	0	0	0

(d) Preference function σ_{gvs}. For all other h', H pairs not specified in the table, $\sigma(h', H, \cdot) = 1$.

Table 3: An example hypothesis class where $\Sigma_{\mathsf{const}}\text{-}\mathsf{TD} = 3$, $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD} = 2$, and $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD} = 1$.

A.2 Proof of Theorem 1

Before we prove our main results for the batch models, we first establish the following results on the non-clashing teaching. The notion of a non-clashing teacher was first introduced by [KKW07]. Our proof is inspired by [KSZ19] which shows the non-clashing property for collusion-free teacher-learner pair, under the batch setting.

Lemma 5 *Assume $\sigma \in \Sigma_{\mathsf{gvs}}$ is collusion-free. Then teacher T must be non-clashing on $\mathcal{H}$. i.e., for any two distinct $h, h' \in \mathcal{H}$ such that $\mathrm{T}(h)$ is consistent with h', $\mathrm{T}(h')$ cannot be consistent with h.*

Proof [Proof of Lemma 5] By definition of the preference function, we have $\forall \sigma \in \Sigma_{\mathsf{gvs}}, h' \in \mathcal{H}$, $\sigma(h'; \mathcal{H}(Z'), \cdot) = g_\sigma(h', \mathcal{H}(Z'))$ for some function g_σ.

We then prove the lemma by contradiction. Assume that the teacher mapping T isn't non-clashing. There exists $h \neq h' \in \mathcal{H}$, where $Z = \mathrm{T}(h)$, and $Z' = \mathrm{T}(h')$ are consistent with both, h and h'.

Assume that the last current hypothesis before the teacher provides the last example of Z is h_1. Then,

$$h = \operatorname*{arg\,min}_{h'' \in \mathcal{H}(Z)} \sigma(h''; \mathcal{H}(Z), h_1) = \operatorname*{arg\,min}_{h'' \in \mathcal{H}(Z \cup Z')} \sigma(h''; \mathcal{H}(Z \cup Z'), h_1) = \operatorname*{arg\,min}_{h'' \in \mathcal{H}(Z \cup Z')} g_\sigma(h'', \mathcal{H}(Z \cup Z')).$$

Where first equality is the definition of a teaching sequence. The second equality is by the definition of collusion-free preference (Definition 1). Similarly we have

$$h' = \operatorname*{arg\,min}_{h'' \in \mathcal{H}(Z' \cup Z)} g_\sigma(h'', \mathcal{H}(Z' \cup Z)).$$

Consequently, $h = h'$, which is a contradiction. This indicates that T is non-clashing. ■

Now we are ready to provide the proof for Theorem 1. We divide the proof of the Theorem 1 into three parts, each corresponding to the equivalence results for a different preference function family.

Proof [Proof of Theorem 1] Part 1 (reduction to TD) and Part 2 (reduction to RTD) of the proof are included in the main paper.

To establish the equivalence between Σ_{gvs}-TD and NCTD, we aim to show that for any hypotheses space $\mathcal{H}$, it holds (i) $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} \geqslant \mathsf{NCTD}(\mathcal{H})$, and (ii) $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} \leqslant \mathsf{NCTD}(\mathcal{H})$.

We first prove (i). According to Lemma 5, for any $\sigma \in \Sigma_{\mathsf{gvs}}$, a successful teacher T with σ is non-clashing on $\mathcal{H}$. Therefore, $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \min_{\text{Successful Teacher T}} \max_{h\in\mathcal{H}} |\mathrm{T}(h)| \geqslant \min_{\text{Non-clashing teacher T}} \max_{h\in\mathcal{H}} |\mathrm{T}(h)| = \mathsf{NCTD}(\mathcal{H})$.

We now proceed to prove (ii). Consider any non-clashing teacher mapping T. First we will prove that there exists $\sigma \in \Sigma_{\mathsf{gvs}}$ such that (T, L_σ) is successful on $\mathcal{H}$. Here L_σ is a learner corresponded to σ as described in §2, and by "successful" we mean that the learner successfully outputs the target hypothesis when teaching terminates. In the following, we construct a preference function σ. First initialize $\sigma(\cdot;\cdot,\cdot) = 1$. Then, for every $h \in \mathcal{H}$, and every S' which $\mathrm{T}(h) \subseteq S'$ and S' is consistent with h assign $\sigma(h;\mathcal{H}(S'),\cdot) = 0$.

We then prove (ii) by contradiction. Consider any set of examples S, and assume there exists two $h' \neq h' \in \mathcal{H}$ where $\sigma(h;\mathcal{H}(S),\cdot) = \sigma(h';\mathcal{H}(S),\cdot) = 0$. Then $\mathrm{T}(h) \subseteq \mathcal{H}(S)$ and $\mathrm{T}(h') \subseteq \mathcal{H}(S)$, also S is consistent with both h and h'. This indicates that, both $\mathrm{T}(h)$ and $\mathrm{T}(h')$ must be consistent with both h, and h'. This contradicts with T being non-clashing. Therefore, for every h, and S' where S' is consistent with h and $\mathrm{T}(h) \subseteq S'$, and $h' \neq h$, we have $\sigma(h;\mathcal{H}(S'),\cdot) < \sigma(h';\mathcal{H}(S'),\cdot)$. Consequently, after providing the examples $\mathrm{T}(h)$ to the learner L_σ, the learner will stay on h even if she receives more consistent labeled examples. Therefore, (T, L_σ) is both collusion-free and successful on $\mathcal{H}$.

Therefore, we conclude that for any teacher mapping T induced by $\sigma \in \Sigma_{\mathsf{gvs}}$, $\max_{h\in\mathcal{H}} |\mathrm{T}(h)| \geqslant \mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma)$. Consequently, $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} \leqslant \mathsf{NCTD}(\mathcal{H})$. Combining this results with (i) hence completes the proof. ■

B Supplementary Materials for §5: Extension of Table 2

This section provides the details of preference functions for the Warmuth class.

$\mathcal{H}$ \ $\mathcal{X}$	x_1	x_2	x_3	x_4	x_5	$\mathcal{S}_{\mathsf{const}} = \mathcal{S}_{\mathsf{global}}$	$\mathcal{S}_{\mathsf{gvs}}$	$\mathcal{S}_{\mathsf{local}}$	$\mathcal{S}_{\mathsf{lvs}}$
h_1	1	1	0	0	0	(x_1, x_2, x_4)	(x_1, x_2)	(x_1)	(x_1)
h_2	0	1	1	0	0	(x_2, x_3, x_5)	(x_2, x_3)	(x_3)	(x_2)
h_3	0	0	1	1	0	(x_1, x_3, x_4)	(x_3, x_4)	(x_3, x_4)	(x_3)
h_4	0	0	0	1	1	(x_2, x_4, x_5)	(x_4, x_5)	(x_5, x_4)	(x_4)
h_5	1	0	0	0	1	(x_1, x_3, x_5)	(x_1, x_5)	(x_5)	(x_5)
h_6	1	1	0	1	0	(x_1, x_2, x_4)	(x_2, x_4)	(x_4)	(x_3)
h_7	0	1	1	0	1	(x_2, x_3, x_5)	(x_3, x_5)	(x_3, x_5)	(x_4)
h_8	1	0	1	1	0	(x_1, x_3, x_4)	(x_1, x_4)	(x_4, x_3)	(x_5)
h_9	0	1	0	1	1	(x_2, x_4, x_5)	(x_2, x_5)	(x_4, x_5)	(x_1)
h_{10}	1	0	1	0	1	(x_1, x_3, x_5)	(x_1, x_3)	(x_5, x_3)	(x_2)

(a) The Warmuth hypothesis class and the corresponding teaching sequences (denoted by $\mathcal{S}$).

h'	$\forall h' \in H$
$\sigma_{\mathsf{const}}(h', \cdot, \cdot)$ $\sigma_{\mathsf{global}}(h', \cdot, \cdot)$	0

(b) σ_{const} and σ_{global}

$h \backslash h'$	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
$\sigma_{\mathsf{local}}(h'; \cdot, h = h_1)$	0	2	4	4	2	1	3	3	3	3
$\cdots$										

(c) σ_{local} representing the Hamming distance between h' and h.

h'	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
H	$\{h_1, h_6\}$	$\{h_2, h_7\}$	$\{h_3, h_8\}$	$\{h_4, h_9\}$	$\{h_5, h_{10}\}$	$\{h_6, h_9\}$	$\{h_7, h_{10}\}$	$\{h_8, h_6\}$	$\{h_9, h_7\}$	$\{h_{10}, h_8\}$
	$\{h_1\}$	$\{h_2\}$	$\{h_3\}$	$\{h_4\}$	$\{h_5\}$	$\{h_6\}$	$\{h_7\}$	$\{h_8\}$	$\{h_9\}$	$\{h_{10}\}$
σ_{gvs}	0	0	0	0	0	0	0	0	0	0

(d) $\sigma_{\mathsf{gvs}}(h'; H, \cdot)$

h'	h_1		h_2		h_3		h_4		h_5	
H	$\{h_1\} \cup \{h_5, h_6, h_8, h_{10}\}^*$		$\{h_2\} \cup \{h_1, h_7, h_6, h_9\}^*$		$\{h_3\} \cup \{h_2, h_7, h_8, h_{10}\}^*$		$\{h_4\} \cup \{h_3, h_6, h_8, h_9\}^*$		$\{h_5\} \cup \{h_4, h_7, h_9, h_{10}\}^*$	
h	h_1		h_1	h_2	h_1	h_3	h_1	h_4	h_1	h_5
σ_{lvs}	0		0	0	0	0	0	0	0	0

$\vdots$

h'	h_6		h_7		h_8		h_9		h_{10}	
H	$\{h_6\} \cup \{h_1, h_4, h_5, h_9\}^*$		$\{h_7\} \cup \{h_1, h_2, h_5, h_{10}\}^*$		$\{h_8\} \cup \{h_1, h_2, h_3, h_6\}^*$		$\{h_9\} \cup \{h_2, h_3, h_4, h_7\}^*$		$\{h_{10}\} \cup \{h_3, h_4, h_5, h_8\}^*$	
h	h_1	h_6	h_1	h_7	h_1	h_8	h_1	h_9	h_1	h_{10}
σ_{lvs}	0	0	0	0	0	0	0	0	0	0

(e) $\sigma_{\mathsf{lvs}}(h'; H, h)$. Here, $\{\cdot\}^*$ denotes all subsets.

Table 4: Teaching sequences with different preference functions for the Warmuth hypothesis class [DFSZ14]

C Supplementary Materials for §5: Proof for Theorem 2

We divide the proof into three parts. The first part shows that the interactions of the two families is Σ_{global}. In part 2 and part 3 of the proof, we show that there exist examples of hypothesis classes, such that $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} > \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}$, or $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} < \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}$.

C.1 Part 1

In this subsection, we provide the full proof for part 1 of Theorem 2, i.e., $\Sigma_{\mathsf{gvs}} \cap \Sigma_{\mathsf{local}} = \Sigma_{\mathsf{global}}$.

Intuitively, observe that the input domains between $\sigma_{\mathsf{local}} \in \Sigma_{\mathsf{global}}$ and $\sigma_{\mathsf{gvs}} \in \Sigma_{\mathsf{gvs}}$ overlaps at the domain of the first argument, which is the one taken by σ_{global}. Therefore, $\forall \sigma \in \Sigma_{\mathsf{global}}, \sigma \in \Sigma_{\mathsf{gvs}} \cap \Sigma_{\mathsf{local}}$. We formalize such idea in the proof below.

Proof Assume $\sigma \in \Sigma_{\mathsf{local}} \cap \Sigma_{\mathsf{gvs}}$. Then, by the definitions of Σ_{local} and Σ_{gvs}, we get

(i) $\exists g^1$, s.t. $\forall h, h' \in \mathcal{H} : \sigma(h'; \cdot, h) = g^1(h', h)$, and
(ii) $\exists g^2$, s.t. $\forall h' \in \mathcal{H}, H \subseteq \mathcal{H} : \sigma(h'; H, \cdot) = g^2(h', H)$

Now consider $h', h^1, h^2 \in \mathcal{H}$, and $H^1, H^2 \subseteq \mathcal{H}$. According to (i), $\sigma(h'; H^1, h^1) = \sigma(h'; H^2, h^1)$. Also, according to (ii) $\sigma(h'; H^2, h^1) = \sigma(h'; H^2, h^2)$. This indicates that, $\forall h', h^1, h^2 \in \mathcal{H}; H^1, H^2 \subseteq \mathcal{H} : \sigma(h'; H^1, h^1) = \sigma(h'; H^2, h^2)$. In other words, there exist $g^3 : \mathcal{H} \to \mathbb{R}$, such that $\forall h' \in \mathcal{H} : \sigma(h'; \cdot, \cdot) = g^3(h')$. Thus, $\sigma \in \Sigma_{\mathsf{global}}$. ■

C.2 Part 2

Part 2. Next, we show that there exists $(\mathcal{H}, \mathcal{X})$, such that $\forall h_0 \in \mathcal{H}$, $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} > \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}$. To prove this statement, we first establish the following lemma.

Lemma 6 *For any $\mathcal{H}$, $\mathcal{X}$, and $h_0 \in \mathcal{H}$, if $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = 1$, then $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = 1$.*

Proof [Proof of Lemma 6] If $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = 1$, there should be some $\sigma_{\mathsf{local}} \in \Sigma_{\mathsf{local}}$, such that $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma_{\mathsf{local}}) = 1$. Now consider σ_{global} such that $\forall h', \sigma_{\mathsf{global}}(h'; \cdot, \cdot) = \sigma_{\mathsf{local}}(h'; \cdot, h_0)$. If $\mathrm{T}_{\sigma_{\mathsf{local}}}$ is the best teacher for σ_{local}, then $\forall h \in \mathcal{H} : |T_{\sigma_{\mathsf{local}}}(h)| = 1$, this indicates that $h = \arg\min_{h' \in \mathcal{H}(T_{\sigma_{\mathsf{local}}}(h))} \sigma_{\mathsf{local}}(h'; \cdot, h_0)$ and $|\arg\min_{h' \in \mathcal{H}(T_{\sigma_{\mathsf{local}}}(h))} \sigma_{\mathsf{local}}(h'; \cdot, h_0)| = 1$. Subsequently, $h = \arg\min_{h' \in \mathcal{H}(T_{\sigma_{\mathsf{local}}}(h))} \sigma_{\mathsf{global}}(h'; \cdot, \cdot)$ and $|\arg\min_{h' \in \mathcal{H}(T_{\sigma_{local}})} \sigma_{\mathsf{global}}(h', \cdot, \cdot)| = 1$. In other words, $\mathrm{T}_{\sigma_{\mathsf{local}}}$ is also a teacher for σ_{local}. This indicates that, $\mathsf{RTD}(\mathcal{H}) = \Sigma_{\mathsf{global}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma_{\mathsf{global}}) = 1$. ■

Now we are ready to provide the proof for Part 2.

Proof [Proof of Part 2 of Theorem 2] We identify $\mathcal{H}$, $\mathcal{X}$, h_0, where $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = 1$ and $\Sigma_{\mathsf{global}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \mathsf{RTD} = 2$. Table 3 illustrates such an example. In the example, since $\mathsf{RTD} = 2$, then by Lemma 6, it must hold that $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} > 1 = \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \mathsf{NCTD}$. ■

C.3 Part 3

Here, we show that there exists a problem instance $(\mathcal{H}, \mathcal{X})$, such that $\forall h_0 \in \mathcal{H}$, $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} < \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}$. Consider the hypothesis class which consists of the powerset $\mathcal{H} = \{0,1\}^k$. First, as proven in Lemma 7 below, we show that $\forall h_0 \in \mathcal{H}$, $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \mathsf{NCTD} \geqslant \lceil k/2 \rceil$.

Lemma 7 (Based on Theorem 23 of [KSZ19]) *Consider the hypothesis class which consists of the powerset $\mathcal{H} = \{0,1\}^k$. Then, $\mathsf{NCTD} \geqslant \lceil k/2 \rceil$.*

Proof First we make the following observation: If T is a non clashing teacher and $h, h' \in \mathcal{H}$ where $h = h' \triangle x$ (i.e., these two hypotheses only differ in their label on one instance), it must be the

case that $(x, h(x)) \in T(h)$, or $(x, h'(x)) \in T(h')$. This holds by nothing that since h, and h' are only different on x, if x is absent in their teaching sequences, this would lead to violation of the non-clashing property of the teacher.

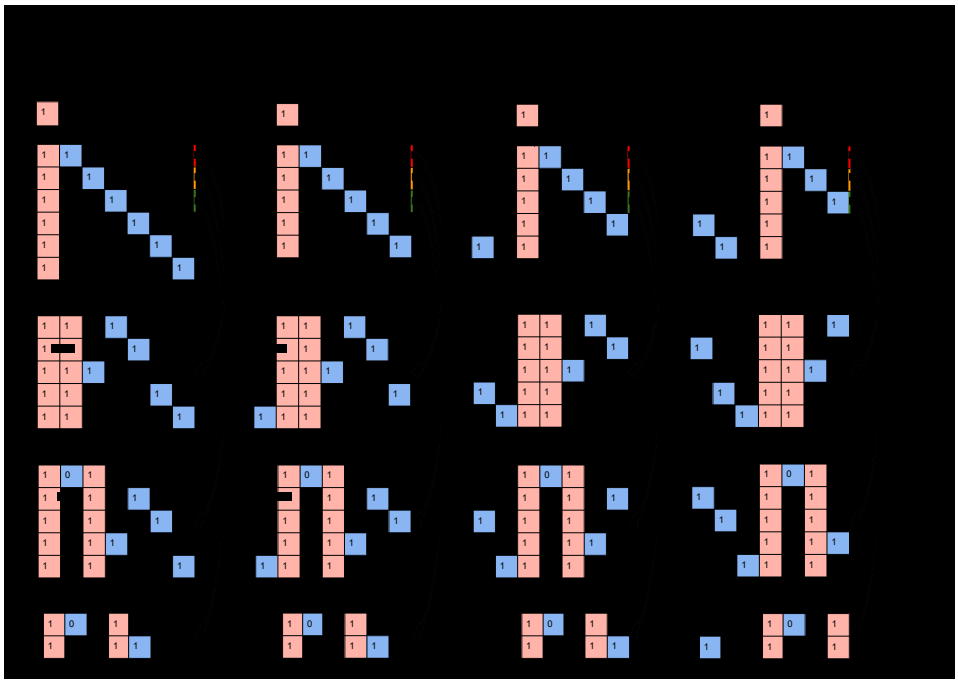
Next we apply this observation on the powerset k hypotheses class where $\mathcal{H}$ consists of all hypotheses which have length k. This indicates that for every $h \in \mathcal{H}$, and $0 \leqslant j \leqslant (k-1)$ all k variants $h \triangle x_j \in \mathcal{H}$. For all $0 \leqslant j \leqslant (k-1)$ by using the above observation, for a pair h and $h \triangle x_j$, we drive $\sum_{i=0}^{2^k-1} |T(h_i)| \geqslant \frac{k \cdot 2^k}{2}$. By applying the pigeon-hole principle, this indicates that there exist an $h \in \mathcal{H}$, where $|T(h)| \geqslant \frac{k}{2}$. In other words $\mathsf{NCTD}(\mathcal{H}) \geqslant \lceil \frac{k}{2} \rceil$. ■

Fix $k = 7$, we get $\Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = \mathsf{NCTD}(\mathcal{H}) \geqslant 4$. On the other hand, we construct a preference function $\sigma \in \Sigma_{\mathsf{local}}$, where $\Sigma_{\mathsf{local}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} \leqslant \mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma) = 3$ for $k = 7$.

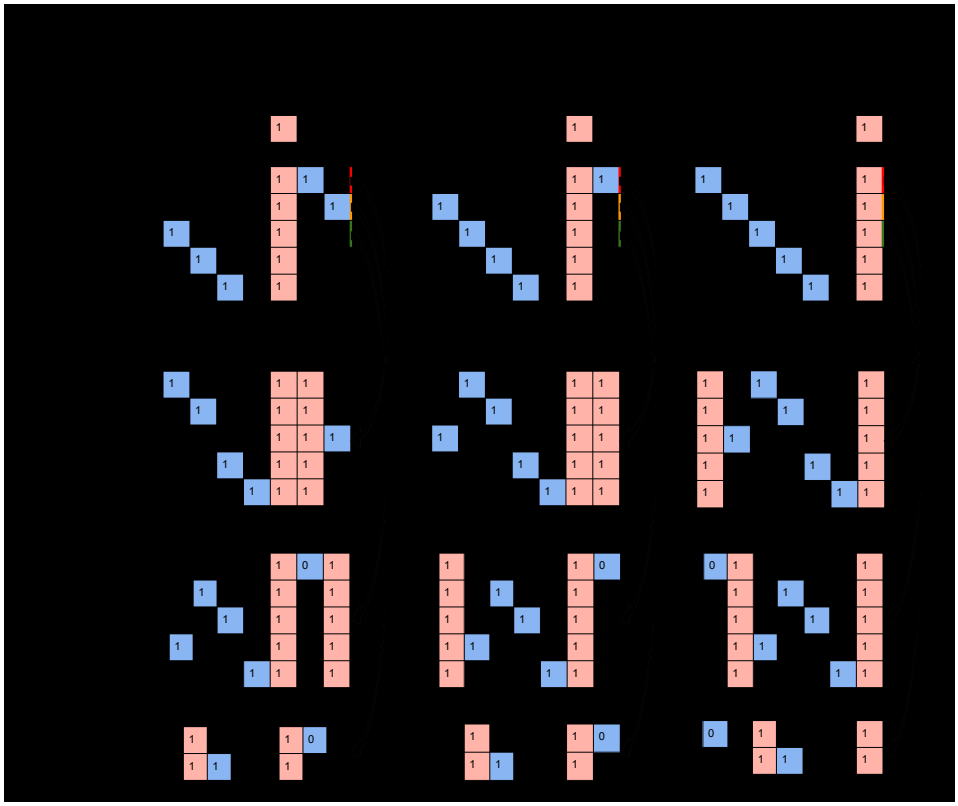
The example is detailed in Figure 5. Intuitively, for any $h_0 \in \mathcal{H}$, we construct a tree of hypotheses with branching factor 7 at the top level, branching factor of 6 at the next level, and so on. Here, each branch corresponds to one teaching example, and each path from h_0 to $h \in \mathcal{H}$ corresponds to a teaching sequence $\mathrm{T}_{\mathsf{local}}(h)$. We need a tree of depth at most 3 to include all the $2^7 = 128$ hypotheses to be taught as nodes in the tree. This gives us a constructive procedure of σ functions achieving $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma) = 3 < \Sigma_{\mathsf{gvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}$, which completes the proof.

h	$\mathcal{X}$	**Preference Function** $\sigma(.;h)$	**Teaching Sequence**
$\underline{h_0}$	0 0 0 0 0 0 0	$h_0 > h_1 > h_2 > h_3 > h_4 > h_5 > h_6 > h_7 >$ others	$((x_0, 0))$
h_1	1 0 0 0 0 0 0	$h_1 > h_8 > h_9 > h_{10} > h_{11} > h_{12} > h_{13}$ others	$((x_0, 1))$
h_8	1 1 0 0 0 0 0	$h_8 > h_{44} > h_{45} > h_{46} > h_{47} > h_{48} >$ others	$((x_0, 1), (x_1, 1))$
h_9	1 1 1 0 0 0 0	$h_9 > h_{79} > h_{80} > h_{81} > h_{82} > h_{83} >$ others	$((x_0, 1), (x_2, 1))$
h_{10}	1 1 1 1 0 0 0	$h_{10} > h_{114} > h_{115} >$ others	$((x_0, 1), (x_3, 1))$
h_{11}	1 1 1 1 1 0 0	$h_{11} >$ others	$((x_0, 1), (x_4, 1))$
h_{12}	1 1 1 1 1 1 0	$h_{12} >$ others	$((x_0, 1), (x_5, 1))$
h_{13}	1 1 1 1 1 1 1	$h_{13} >$ others	$((x_0, 1), (x_6, 1))$
h_{44}	1 1 0 1 0 0 0	$h_{44} >$ others	$((x_0, 1), (x_1, 1), (x_3, 1))$
h_{45}	1 1 0 1 1 0 0	$h_{45} >$ others	$((x_0, 1), (x_1, 1), (x_4, 1))$
h_{46}	1 1 1 0 1 0 0	$h_{46} >$ others	$((x_0, 1), (x_1, 1), (x_2, 1))$
h_{47}	1 1 0 0 0 1 0	$h_{47} >$ others	$((x_0, 1), (x_1, 1), (x_5, 1))$
h_{48}	1 1 0 0 1 0 1	$h_{48} >$ others	$((x_0, 1), (x_1, 1), (x_6, 1))$
h_{79}	1 0 1 0 0 0 0	$h_{79} >$ others	$((x_0, 1), (x_2, 1), (x_1, 0))$
h_{80}	1 0 1 0 1 0 0	$h_{80} >$ others	$((x_0, 1), (x_2, 1), (x_4, 1))$
h_{81}	1 0 1 0 1 1 0	$h_{81} >$ others	$((x_0, 1), (x_2, 1), (x_5, 1))$
h_{82}	1 1 1 1 0 1 0	$h_{47} >$ others	$((x_0, 1), (x_2, 1), (x_3, 1))$
h_{83}	1 0 1 1 1 0 1	$h_{83} >$ others	$((x_0, 1), (x_2, 1), (x_6, 1))$
h_{114}	1 0 0 1 0 0 0	$h_{114} >$ others	$((x_0, 1), (x_3, 1), (x_1, 0))$
h_{115}	1 0 0 1 1 0 0	$h_{115} >$ others	$((x_0, 1), (x_3, 1), (x_4, 1))$

Table 5: More details about Figure 5: This table lists down all the hypotheses in the left branch of the tree. For each of these hypotheses, it shows the preference function from the hypothesis, as well as the teaching sequence to teach the hypothesis. Consider h_9: We have $\sigma(., h_9) = \{h_9 > h_{79} > h_{80} > h_{81} > h_{82} > h_{83} >$ others$\}$. Also, we have teaching sequence for h_9 as $\{(x_0, 1), (x_2, 1)\}$.



(a) This figure is representing the teaching sequence for first four for direct children of h_0 (top four most preferred hypothesis of h_0 after h_0) and all of their children.



(b) This figure is representing the teaching sequence for next three direct children of h_0 (next three most preferred hypothesis of h_0) and all of their children.

Figure 5: Details of teaching sequences, for a preference function $\sigma \in \Sigma_{\mathsf{local}}$, where $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma) = 3$ for powerset $k = 7$ class. For any hypothesis the cell with blue color is representing last teaching example in teaching sequence, and the cells with red color are representing rest of teaching sequence. Also see Table 5 that lists down details for all the hypotheses in the left branch of the tree.

D Supplementary Materials for §5.3

D.1 Proof of Lemma 4

In this section, we extend the proof sketch of Lemma 4 in the main paper into the full proof. A useful notion for this proof is the notion of *H-distinguishable set*:

Definition 7 (Based on [DFSZ14]) *A set of instances $X \subseteq \mathcal{X}$ is H-distinguishable, if $|H_{|X}| = |H|$.*

For completeness, we also incorporate part of the proof sketch from §5.3 into the extended proof below.

Proof [(Extended) Proof of Lemma 4] Let us define $H_x = \{h \in H : h \triangle x_{|\Psi_H} \in H_{|\Psi_H}\}$. Here, $h \triangle x$ denotes the hypothesis that only differs from h on the label of x. Fix a reference hypothesis h_H. $\forall x_j \in \Psi_H$, 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 $H^{y_1}_{x_1} = \{h \in H_{x_1} : h(x_1) = y_1\}$ as the first partition.

H \ Ψ_H	x_1	x_2	...	x_{m-1}
h_0	0	00 ... 0		
h_1	0	a		
h_2	1	b		
h_3	0	b		
h_4	1	c		
h_5	1	d		
h_6	1	e		
h_7	1	a		

(a) H

H_{x_1} \ Ψ_H	x_1	x_2	...	x_{m-1}
h_1	0	a		
h_2	1	b		
h_3	0	b		
h_7	1	a		

(b) H_{x_1}

H^1 \ Ψ_H	x_1	x_2	...	x_{m-1}
h_2	1	b		
h_7	1	a		

(c) $H^1 = H^{y_1=1}_{x_1}$

Table 6: Illustrative example for constructing the first partition $H^1 = H^{y_1=1}_{x_1}$.

In Table 6, we provide an example hypothesis class where we show how to construct the first partition $H^{y_1}_{x_1}$. Table 6a shows the hypothesis class $\mathcal{H}$ (here $a \neq b \neq c \neq d \neq e$) and $h_{\mathcal{H}} = h_0$. Table 6b shows the resulting set of hypotheses $H_{x_1} = \{h \in H : h \triangle x_{1|\Psi_H} \in H_{|\Psi_H}\}$, and Table 6c shows the first partition $H^{y_1=1}_{x_1}$.

We denote $H^1 := H^{y_1}_{x_1}$, and define $\Psi_{H^1} \subseteq \Psi_H \backslash \{x_1\}$ to be any compact-distinguishable set on H^1.

Lower VCD. Let $d = \mathsf{VCD}(H, \Psi_H)$. In the following, we prove that $\mathsf{VCD}(H^1, \Psi_{H^1}) \leqslant d - 1$. We consider the following two cases:

1. If $d > 1$, then
   $\mathsf{VCD}(H^1, \Psi_{H^1}) \leqslant \mathsf{VCD}(H^{y_1}_{x_1}, \Psi_H) = \mathsf{VCD}(H_{x_1}, \Psi_H) - 1 \leqslant \mathsf{VCD}(H, \Psi_H) - 1 \leqslant d - 1$
   Since $\Psi_{H^1} \subset \Psi_H$, the first inequality is due to the monotonicity of VCD. The equality follows from the fact that, for all $h \in H^{y_1}_{x_1}$, it holds that $h(x_1) = y_1$ and $h \triangle x_{1|\Psi_H} \in H_{x_1|\Psi_H}$. This indicates that, $X \subseteq \Psi_H$ shatters $H^{y_1}_{x_1}$, iff $X \cup \{x_1\}$ shatters H_{x_1}. The second inequality comes from the fact that VCD is monotonic.
2. If $d = 1$ and $|H^{y_1}_{x_1}| \geqslant 2$, then
   similar to the previous case we have the following: $\mathsf{VCD}(H_{x_1}, \Psi_H) \leqslant \mathsf{VCD}(H, \Psi_H) = 1$ and $\mathsf{VCD}(H_{x_1}, \Psi_H) = \mathsf{VCD}(H^{y_1}_{x_1}, \Psi_H) + 1$. Subsequently, $\mathsf{VCD}(H^1, \Psi_{H^1}) = 0$.
3. If $d = 1$ and $|H^{y_1}_{x_1}| < 2$, then
   since $|H^{y_1}_{x_1}| < 2$, by definition, we have $\mathsf{VCD}(H^1, \Psi_{H^1}) = 0$ and hence is less than $d = 1$.

That is, the first partition H^1, Ψ_{H^1} has $\mathsf{VCD}(H^1, \Psi_{H^1}) \leqslant d - 1$, i.e., H^1 satisfies property (i). In addition, it is clear that $H^1{}_{|\{x_1\}} = \{y_1\} = \{1 - h_H(x_1)\}$. Therefore, H^1 also satisfies property (ii).

Non-emptiness of H^1. For the sake of contradiction assume that H^1 is empty. Note that Ψ_H is H-distinguishable. Since H^1 is empty, this means that there is no pair of hypotheses that differ only on x_1. This in turn indicates that $\Psi_H \backslash \{x_1\}$ is H-distinguishable. However, $|\Psi_H \backslash \{x_1\}| < |\Psi_H|$ and this is in contradiction to the assumption that Ψ_H is compact-distinguishable on H.

Continuing to create partitions. Next, we remove the first partition H^1 from H, and continue to create the above mentioned partitions on $H_{\text{rest}} = H \backslash H^1$ and $X_{\text{rest}} = \Psi_H \backslash \{x_1\}$. We claim that $H_{\text{rest}}, X_{\text{rest}}$ exhibit the following properties.

1. X_{rest} is H_{rest}-distinguishable (see Definition 7).

   For the sake of contradiction, assume that there exists a pair of hypotheses $h^1, h^2 \in H_{\text{rest}}$ such that $h^1_{|X_{\text{rest}}} = h^2_{|X_{\text{rest}}}$. However, we know that $h^1_{|\Psi_H} \neq h^2_{|\Psi_H}$. Then, these two hypotheses should have been in H_{x_1} and only one of them could have stayed in H_{rest}. Hence, there is no such pair of hypotheses in H_{rest} and this completes the proof of the statement.

2. X_{rest} is also a compact-distinguishable on H_{rest}.

   We now provide a concrete proof for the above statement. Imagine $X \subseteq X_{\text{rest}}$ is an H_{rest}-distinguishable set. In the following, we prove that $X \cup \{x_1\}$ is H-distinguishable.

   For the sake of contradiction assume that, $X \cup \{x_1\}$ isn't H-distinguishable. This indicates that there exist two hypotheses $h^1 \neq h^2 \in H$, where they are equal on $X \cup \{x_1\}$, i.e., $h^1_{|X\cup\{x_1\}} = h^2_{|X\cup\{x_1\}}$; also this implies $h^1_{|X} = h^2_{|X}$. Since $H = H_{\text{rest}} \cup H^1$, we consider the following three cases.

   (i) $h^1, h^2 \in H_{\text{rest}}$. Since X is H_{rest}-distinguishable, it is a contradiction that $h^1_{|X} = h^2_{|X}$.

   (ii) $h^1, h^2 \in H^1$. By the construction of H^1, there exist $\hat{h}^1, \hat{h}^2 \in H_{\text{rest}}$, such that $\hat{h}^1_{|X\cup\{x_1\}} = h^1 \triangle x_{1|X\cup\{x_1\}}$ and $\hat{h}^2_{|X\cup\{x_1\}} = h^2 \triangle x_{1|X\cup\{x_1\}}$. Furthermore, since $h^1_{|X} = h^2_{|X}$, we must have $\hat{h}^1_{|X} = \hat{h}^2_{|X}$, which contradicts the fact that X is H_{rest}-distinguishable.

   (iii) $h^1 \in H^1, h^2 \in H_{\text{rest}}$. By the construction of H^1, there exist $\hat{h}^1 \in H_{\text{rest}}$, such that $\hat{h}^1_{|X\cup\{x_1\}} = h^1 \triangle x_{1|X\cup\{x_1\}}$. Furthermore, since $h^1_{|X} = h^2_{|X}$, we must have $\hat{h}^1_{|X} = h^2_{|X}$, which contradicts the fact that X is H_{rest}-distinguishable.

   Therefore, we conclude that $X \cup \{x_1\}$ is H-distinguishable. Recall that Ψ_H is compact-distinguishable on H. This indicates that $\Psi_H = X \cup \{x_1\}$, and subsequently $X = X_{\text{rest}}$. This indicates that X_{rest} is compact-distinguishable on H_{rest}.

3. If $|\Psi_H| > 1$, then $|H_{\text{rest}}| > 1$.

   We prove the above statement by contradiction. Assume that $|H_{\text{rest}}| = 1$. Since we know that H^1 is non empty, hence $|H_{\text{rest}}| = 1$ implies that $|H^1| = 1$. Let $H^1 = \{h\}$, and $H_{\text{rest}} = \{h'\}$, then $h'_{|\Psi_H} = h \triangle x_{1|\Psi_H}$. Since we know that $H = H^1 \cup H_{\text{rest}}$, subsequently $\{x_1\}$ is compact-distinguishable on H, which is in contradiction to the assumption that Ψ_H is compact-distinguishable.

Case of $|X_{\text{rest}}| > 1$. 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_{rest} reduces to 1, i.e. $X_{\text{rest}} = \{x_{m-1}\}$. Until then, we obtain partitions $\{H^1, ..., H^{m-2}\}$. By construction, H^j satisfy properties (i) and (ii) for all $j \in [m-2]$.

Note that each step X_{rest} is compact H_{rest}-distinguishable set. This implies that we have never lost a hypothesis in this process, i.e., all hypotheses in H were either in one of H_j's or in H_{rest}.

Case of $|X_{\text{rest}}| = 1$. It remains to show that the last two partitions H^{m-1} and H^m also satisfy properties (i) and (ii); additionally we need to satisfy property (iii). Since $X_{\text{rest}} = \{x_{m-1}\}$, and $|H_{\text{rest}}| > 1$ before we start iteration $m-1$, there must exist exactly two hypotheses in H_{rest}. Therefore $|H^{m-1}|, |H^m| = 1$, and $H^{m-1}_{|\{x_{m-1}\}} = \{\{y_{m-1}\}\}$. This implies that $\mathsf{VCD}(H^{m-1}, \Psi_{H^{m-1}}) = \mathsf{VCD}(H^m, \Psi_{H^m}) = 0 \leqslant d-1$. Furthermore, notice that for every $j \in [m-1], h \in H^j, h_H(x_j) \neq h(x_j)$. This indicates $h_H \in H_m$. Since $|H_m| = 1$, we get $H_m = \{h_H\}$ which completes the proof. ■

D.2 Supplementary Materials for the Proof of Theorem 3

Our proof of Theorem 3 in the main paper relies on the fact that every teaching example (x_j, y_j), where $x_j \in \Psi_H$ and $y_j = 1 - h(x_j)$ for some fixed h, corresponds to a unique version space V^j. The proof depends on the following lemma.

Lemma 8 *Fix $H \subseteq \mathcal{H}$, and let $\Psi_H \subseteq \mathcal{X}$ be a compact-distinguishable set on H. For any $x, x' \in \Psi_H$ and $y, y' \in \{0, 1\}$ such that $(x, y) \neq (x', y')$, the resulting version spaces $\{h \in H : h(x) = y\}$ and $\{h \in H : h(x') = y'\}$ are different.*

Proof [Proof of Lemma 8] Denote $A = \{h \in H : h(x) = y\}$ and $B = \{h \in H : h(x') = y'\}$. We consider the following two cases: (1) $y = y'$, and (2) $y \neq y'$. For the first case where $y = y'$, if $A = B$, this would violate the first part of property (i) of Lemma 4, (i.e., there do not exist distinct x, x' s.t. $\forall h \in H : h(x) = h(x')$. For the second case, if $A = B$, this would violate the second part of property (i). Hence it completes the proof. ■