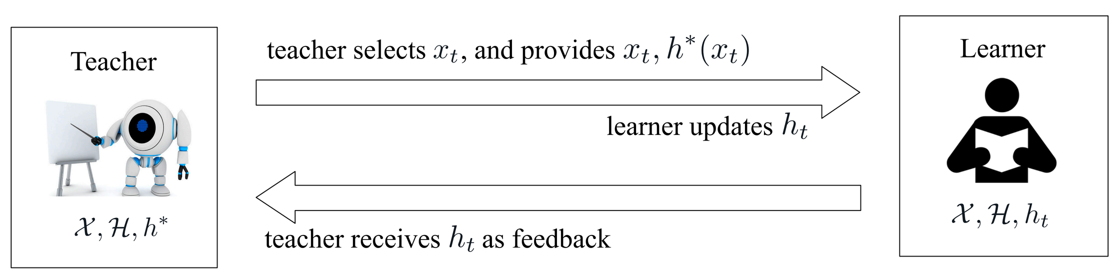
PREFERENCE-BASED BATCH AND SEQUENTIAL TEACHING: TOWARDS A UNIFIED VIEW OF MODELS

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

†Max Planck Institute for Software Systems (MPI-SWS) ‡University of Chicago ⋆University of Wisconsin-Madison

Algorithmic Teaching



Canonical Example

Discrete points on a line

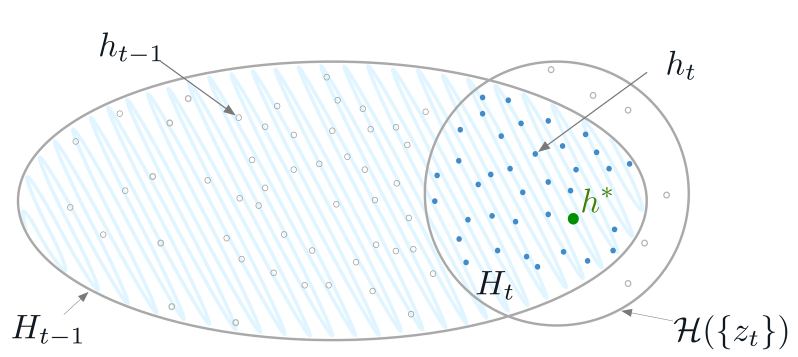
$\mathcal{X} = \{1, 2, \dots, n\}$

Threshold classifier $h(x) = 1$ iff $x \geq \theta_h$ where $\theta_h \in \{1, 2, \dots, n\}$

Complexity of passive learning: $O(n)$; active learning: $O(\log(n))$; teaching: 2.

Interaction Protocol

1: learner's initial version space is $H_0 = \mathcal{H}$ and learner starts from $h_0 \in \mathcal{H}$
2: **for** $t = 1, 2, 3, \dots$ **do**
3: teacher provides $z_t = (x_t, h^*(x_t))$
4: learner updates $H_t = H_{t-1} \cap \mathcal{H}(\{z_t\})$; picks $h_t \in H_t$
5: teacher receives h_t as feedback from the learner
6: **if** $h_t = h^*$ **then** teaching process terminates



Complexity Measures

Notions	Description
TD	classical worst-case teaching complexity
RTD	notion of TD when teaching a collaborative learner
NCTD	strongest notion of TD that respects collusion-freeness
Local-PBTD	teaching complexity of a weak sequential model

Research Questions

- Is there a framework unifying different notions of TD's?
- Can we identify models with teaching complexity linear in the Vapnik–Chervonenkis dimension VCD?

Our Contributions

A novel framework capturing the teaching process via preference functions Σ, where each function $\sigma \in \Sigma$ induces a teacher-learner pair. Our main results are as follows:

- We show that existing batch models correspond to specific families of σ functions in our framework.
- We identify sequential models with teaching complexity linear in the VCD of the hypothesis class.
- We provide a constructive procedure to find σ functions with low teaching complexity.

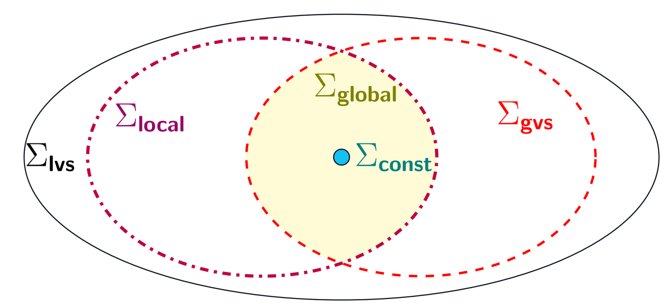


Table 1: Main Resuls

Families	Σ_{const}	Σ_{global}	Σ_{gvs}	Σ_{local}	Σ_{lvs}
Reduction	TD	RTD	NCTD	Local-PBTD	–
Complexity Results	–	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD}^2)$	$O(\mathsf{VCD})$
	[GK95]	[Zil+11]	[KSZ19]	[Che+18]	–

Learner's Preference Function

A preference function $\sigma : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R}$ models how a learner navigates in the version space as it receives teaching examples (*line 4 of Interaction Protocol*):

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

Teaching Complexity Σ-TD

Teaching Dimension for a Preference Function

Fix $\mathcal{X}$, $\mathcal{H}$, and learner's initial hypothesis h_0. Let $D_{\mathcal{X},\mathcal{H},h_0}(\sigma, h^\star)$ be the worst-case optimal cost for steering the learner from h_0 to $h^\star$ for some preference function σ. Then, the teaching dimension w.r.t. σ is defined as the worst-case optimal cost for teaching any target $h^\star$:

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

Teaching Dimension for a Family of Preference Functions

We define the teaching dimension for a family Σ as the teaching dimension w.r.t. the best $\sigma \in \Sigma$:

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

Collusion-free Preference Functions

Definition 1 (Collusion-free teaching [GM96] (batch setting)) *A learner outputting hypothesis h will not change its output if given additional information consistent with h.*

Definition 2 (Collusion-free preference (this paper)) *If h is the only hypothesis in the most preferred set defined by σ, then the learner will stay at h if additional information received by the learner is consistent with h.*

We study preference functions that are collusion-free as per Definition 2:

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

Preference-based Teaching Models

- Batch models:
  - $\Sigma_{\mathsf{const}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists c \in \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = c\}$
  - $\Sigma_{\mathsf{global}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists\, g : \mathcal{H} \to \mathbb{R}, \text{ s.t. } \forall h', H, h,\ \sigma(h'; H, h) = g(h')\}$
  - $\Sigma_{\mathsf{gvs}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists\, g : \mathcal{H} \times 2^{\mathcal{H}} \to \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = g(h', H)\}$
- Sequential models:
  - $\Sigma_{\mathsf{local}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists\, g : \mathcal{H} \times \mathcal{H} \to \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = g(h', h)\}$
  - $\Sigma_{\mathsf{lvs}} = \{\sigma \in \Sigma_{\mathsf{CF}} \mid \exists\, g : \mathcal{H} \times 2^{\mathcal{H}} \times \mathcal{H} \to \mathbb{R}, \text{ s.t. } \forall h', H, h, \sigma(h'; H, h) = g(h', H, h)\}$
- Teaching sequences with different preference functions for the Warmuth hypothesis class:

$h \backslash x$	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
$\dots$										

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

Main Results

- Reduction to existing notions of TD's (see Table 1).
- Proving $\Sigma_{\mathsf{lvs}}\text{-}\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0} = O\big(\mathsf{VCD}(\mathcal{H}, \mathcal{X})\big)$ via a constructive procedure.

Key Ideas for Constructing $\sigma \in \Sigma_{\mathsf{lvs}}$ with $\mathsf{TD}_{\mathcal{X},\mathcal{H},h_0}(\sigma) = O(\mathsf{VCD})$

- Introducing a new notion of *compact distinguishable set.*
- Partitioning the hypothesis class into subsets of hypothesis classes with lower VCD using the compact distinguishable set.
- Recursively applying the partitioning procedure to create the preference function σ.

Discussions

- Designing σ functions for addressing the open question of whether RTD is linear in VCD.
- Designing teaching algorithms for sequential models.

References

[GK95] Sally A Goldman and Michael J Kearns. "On the complexity of teaching". In: *J. Comput. Syst. Sci* 50.1 (1995), pp. 20–31.

[Zil+11] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. "Models of cooperative teaching and learning". In: *JMLR* 12.Feb (2011), pp. 349–384.

[KSZ19] David Kirkpatrick, Hans U. Simon, and Sandra Zilles. "Optimal Collusion-Free Teaching". In: *ALT.* Vol. 98. 2019, pp. 506–528.

[Che+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: *NeurIPS*. 2018, pp. 1476–1486.

[GM96] Sally A Goldman and H David Mathias. "Teaching a smarter learner". In: *J. Comput. Syst. Sci* 52.2 (1996), pp. 255–267.