A New Probabilistic Machine Teaching Model That Redefines Cheating ⋆

César Ferri[a], José Hernández-Orallo[a], Jan Arne Telle[b]

[a] *VRAIN, Universitat Politècnica de València, Spain*
[b] *Department of Informatics, University of Bergen, Norway*

Abstract

Machine teaching has developed several models in which a teacher sends a witness, usually in the form of a set of examples (ground elements), and the learner must identify the concept the teacher has in mind. Since the teacher and learner can collude by assigning each witness with any desired concept regardless of everything else, several restrictions have been introduced to avoid 'cheating', such as collusion-free or non-clashing teaching. However, these restrictions forbid several teaching situations that we intuitively consider natural and fair, especially those 'changes of mind' of the learner as more evidence is given, affecting the likelihood of concepts and ultimately their posteriors. Under a new generalised probabilistic teaching, not only do these non-cheating constraints look too narrow but we also show that the most relevant machine teaching models are particular cases of this framework: the consistency graph between concepts and elements simply becomes a joint probability distribution. We show a standard procedure, called WSC (Witness Sampling Composition), that from a joint distribution on ground elements builds the witness joint distribution (over sets of ground elements). Using this WSC, if teacher and learner share a ground joint distribution corresponding to their beliefs about the factual world, choosing a unique concept with maximum posterior over the witness joint distribution is *not* cheating. We prove a chain of relations, also with a theoretical lower bound, on the teaching dimension of the old and new models. Overall, this new setting is more general than the traditional machine teaching models, yet at the same time more intuitively

⋆An extended abstract of this paper was presented at the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022) [5].

Email addresses: `cferri@dsic.upv.es` (César Ferri), `jorallo@dsic.upv.es` (José Hernández-Orallo), `Jan.Arne.Telle@uib.no` (Jan Arne Telle)

capturing a less abrupt notion of non-cheating teaching.

---

1. Introduction

What is the most efficient way for a teacher to make a learner identify a given concept? This question, when addressed from an algorithmic perspective, takes the name of 'machine teaching' [21], although teacher and learner may be computer models or humans. A very common framework for teaching consists of a set of examples and a set of concepts over those examples. Unbeknownst to the learner, the teacher picks a concept and, by showing a subset of examples, or *witness*, makes the learner single out the concept the teacher has in mind. An important complexity notion is the teaching dimension (TD) of a concept class C, which is the minimum number of examples, from a set of ground elements X, needed to teach any concept in the class. As the teaching complexity depends on the protocol between teacher and learner and their shared information, different teaching models lead to different values for the teaching dimension.

Over the past decades, several teaching models have been proposed, for example, the classical teaching (CT) model [8], the optimal teacher (OT) model [1], recursive teaching (RT) [23, 2, 3], preference-based teaching (PBT) [7, 6], and non-clashing teaching (NCT) [12]. In all these models, the teacher $T : C \rightarrow W$ is viewed as a mapping from concepts $c \in C$ to witness $w \in W$ (usually sets of positively labelled examples from X) and the learner $L : W \rightarrow C$ as a *partial* mapping in the opposite direction. Moreover, the examples $T(c)$ employed to teach concept c must be consistent with c, and the guessed concept $L(w)$ when given example set w must also be consistent with w. A successful teacher-learner pair has $L(T(c)) = c$ for any concept in the class.

But there are ways in which a teacher and a learner can fix the game beforehand. This is known as cheating, or unfair collusion between teacher and learner. As described by Moran et al [15], "roughly speaking, a collusion occurs when teacher and student agree in advance on some unnatural encoding of information about the concept c using the bit description of the chosen examples, instead of using attributes that separate c from the other concepts". Goldman and Mathias [1996] proposed that a model should be called collusion-free if whenever $T(c) \subseteq w$ and w is consistent with c, denoted by $c \models w$, then also $L(w) = L(T(c)) = c$ (hereafter called GM-collusion-free). Many abstract teaching models in the literature were introduced specifically to improve the teaching complexity of previous models while remaining GM-collusion-free. For example, the five models mentioned above (CT, OT, RT, PBT and NCT), in this order, have strictly improving teaching complexities, and all remain GM-collusion-free. The non-clashing model is provably the end of this line, as it can be shown that if every concept in class C can be taught with at most k examples by some GM-collusion-free model then the same holds for the non-clashing model, since the non-clashing model actually adheres to no other constraints than those formulated by the Goldman and Mathias condition.

However, in a GM-collusion-free model a learner guessing c is not allowed to change its mind if given additional examples consistent with c. Consider a learner that sees the witness $w = \{3\}$, composed of one single ground element $3 \in X = \mathbb{N}$, and assigns some plausibility to the hypothesis that the underlying concept is the set of odd numbers c_{odd}. Some other plausibility is given to other hypotheses, such as the powers of three c_{pow3}, the prime numbers c_{prime}, etc. Based on simplicity of consistent concepts, the learner guesses c_{odd}. Now, if the same learner sees the witness $\{3, 29\}$ or $\{3, 11\}$, the powers of three is ruled out. But the *likelihood* of these examples for c_{prime} now looks higher, even higher than for c_{odd}, so that the learner now guesses c_{prime}. Adding more examples consistent with a concept (initially guessed as the odd numbers) may end up in a change of the guess (to the prime numbers), in a very natural way, while forbidden in all GM-collusion-free models.

We claim that all this is more naturally understood by extending the notion of the consistency graph between concepts and witness into a *witness joint distribution* $p : C \times W \rightarrow [0, 1]$. Both teacher and learner share $p(c, w)$ for every pair of concept and witness, with $p(c, w) > 0$ if and only if $c \models w$. In this framework, the learner L is just defined as choosing the concept that uniquely maximises the posterior $L(w) = \arg!\max_c p(c|w)$, which can be calculated from the witness joint distribution and its marginals as $p(c, w)/p(w)$, whenever a particular w is given by the teacher. With this framework we clearly see that the CT model simply assumes $p(c, w)$ such that $p(w|c)$ and $p(c)$ are both uniform, while the PBT model allows for non-uniform concept priors $p(c)$. However, we get the new maximum likelihood (MLE) teaching model, where $p(w|c)$ is free but $p(c)$ is uniform, and the most general case, the maximum a posteriori (MAP) teaching model, where all probabilities are chosen freely provided they make up a valid joint distribution $p(c, w)$.

If the learner does derive its posterior from $p(c, w)$, should $p(c, w)$ be defined in any natural way? In the beginning, teacher and learner share a set of ground elements X, from which the whole teaching process is built: a witness is a new structure that is composed in different ways depending on the teaching paradigm. One common way of building the set of witness objects is simply $W = 2^X$, i.e., a witness is a set of ground elements. But we can also have negative examples with $W = 2^{X \times \{-,+\}}$. These two cases represent a situation where the witness is built by composing elements from X without replacement. But we can also build witnesses with replacement, as when $W = X^*$, with X^* being the set of all finite sequences that can be built from X. Under this perspective, we see that the *witness* joint probability $p(c, w)$ *should* derive from a more fundamental distribution, the *ground* joint distribution $q(c, x)$, defined as $q : C \times X \to [0, 1]$, as an extension of the consistency graph between concepts and ground elements.

We claim that a natural setting for non-cheating teaching must be based on teacher and learner only sharing q, the joint distribution on ground elements and concepts. We claim that whatever this q is, if it corresponds to the beliefs teacher and learner have about the factual world, then there is no cheating. From here, witnesses can be constructed by composing these ground elements in different ways, e.g., sets of positive examples, multisets of positive examples, sets of positive and negative examples, or other structures. In this paper we focus on the first two, sets and multisets. We present a unifying way of deriving p from q in these situations, which is based on the notion of Witness Sampling Composition (WSC), where the joint distribution of concepts and witnesses is derived by composing the witnesses by sampling from the ground elements, with or without replacement depending on the case of multisets or sets. We postulate that this model intuitively matches the notion of non-cheating teaching.

The main contributions of this paper are:

- We show that the use of witness joint distribution is a unifying framework, by fleshing out that several teaching paradigms can be expressed by different constraints on priors and likelihoods (Table 1). The GM-collusion property and a probabilistic version of it known as monotonocity hold when the likelihood is uniform.

- For the two new machine teaching paradigms in Table 1, MLE and MAP, we show that monotonicity and GM-collusion do not hold (and are not equivalent).

- We propose a new notion of non-cheating machine teaching, where we argue that T and L can share *any factual* joint distribution on ground elements q. Assuming WSC we derive the witness joint distribution p for witness sets and multisets by applying a composition of probabilities as a sampling process from the joint distribution without and with replacement respectively.

$$\boxed{\begin{array}{c} LBTD^{++} = JDTD^{++} < STD^{++} \leq \\ \leq LBTD^{+} = JDTD^{+} < STD^{+} \leq NCTD^{+} \end{array}}$$

Figure 1: Summary of relationships shown between teaching dimensions of the old and new machine teaching models.

- We show that the theoretically lowest-bound TD for sets ($LBTD^+$) and multisets ($LBTD^{++}$) of positive examples can be achieved by some witness joint distribution ($JDTD$). The new WSC machine teaching model, yielding the Sampling Teaching Dimension STD, is less powerful than $JDTD$, but more powerful than the Non-Clashing TD ($NCTD$). The precise chain is shown in Figure 1. In sum, we argue that the Sampling Teaching Dimension STD allows for multiple changes of mind, and can achieve lower TD than other classical teaching models, while being non-cheating under our new setting whenever the ground distribution used corresponds to factual beliefs.
- We show a polynomial-time algorithm to compute $LBTD^+(C) = JDTD^+(C)$, for any concept class C over a domain X where $|C|$ is exponential in $|X|$, and leave as an open problem if polynomial-time solvability can be achieved also in other cases.

The new generalised probabilistic framework presented in this paper reconnects the traditional notions of machine teaching with modern probabilistic views of machine teaching (including Bayesian teaching), and gives a completely different perspective on what teacher and learner should be allowed to share and how they should derive their choices according to this shared information.

2. Cheating and Probabilities in MT

In the classical model for machine teaching of Goldman and Kearns (1995) the shared information between teacher and learner consists of whether every example is consistent with a concept or not. Note that this information is between the ground elements in X, and the concepts in C. Then, a witness set can be built in different ways. When only positive examples are allowed, $W = 2^X$ and this consistency information is extended from X to W. If c is consistent with x_1 and x_2 then it has to be consistent with $\{x_1, x_2\}$. We will call this compatibility relation the *consistency graph* and view it as a bipartite graph between concepts

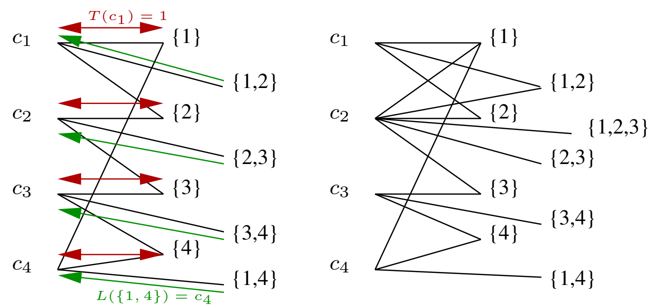

Figure 2: Consistency graph with $|C| = |X| = 4$ and witness sets $W = 2^X$. No concept consistent with a witness of size 3 or 4. Teacher mapping in red. Learner mapping in red and green. $NCTD = 1$, $CTD = 2$, $PBTD = 2$. Right: Adding $c_2 \models 1$ increases $NCTD$.

and witnesses with adjacency denoting consistency, as done by Kirkpatrick et al (2020). A witness w will then uniquely identify a concept c if the only edge incident to w is cw.

This demand seemed too strict, and other models where lower complexity could be achieved were considered. However, it became paramount to avoid cheating. In 1996 Goldman and Mathias proposed that a model was collusion-free if further consistent evidence did not make the learner change its mind, i.e., if $L(T(c)) = c$ then for any superset w' of $T(c)$, if cw' is an edge of the consistency graph then $L(w') = c$.

Consider Figure 2 (left); under the classical setting where the only shared information is the consistency graph on black edges, we have minimal GM-collusion-free teaching dimension ($NCTD = 1$) by the teacher function corresponding to the red matching saturating the concepts. However, TD=1 cannot be achieved with a learner based on concept preferences only, as there is a long cycle for the singletons and we thus have $CTD = PBTD = 2$. Note that if we add $c_2 \models 1$ and the edges $c_2\{1\}$ and $c_2\{1,2\}$ and $c_2\{1,2,3\}$ to the consistency graph (Figure 2 right) then $NCTD = 2$. In particular, the red matching is no longer GM-collusion-free, as $w = \{1,2\}$ is a superset of both $T(c_1)$ and $T(c_2)$.

Kirpatrick et al in [2019] proved a theorem stating that a teacher function allows for GM-collusion-free teaching if and only if the consistency graph does not have any induced cycle on 4 edges with 2 of them being chosen by the teacher (as would happen in Figure 2 right, if the red edges in left are chosen by the teacher). They called such a teaching protocol non-clashing and one could view it, in retrospect, as an alternative definition for a GM-collusion-free model of teaching.

In preference-based teaching (PBT) the unique identification rule is relaxed so that the learner will identify c from w, i.e. $L(w) = c$, as long as w has no edge in the consistency graph to any concept c' with higher preference than c. It is easy to see that for any teacher and learner adhering to this rule the resulting protocol is GM-collusion-free, as for any node w' representing a superset of w, the neighbours of w' will be a subset of the neighbours of w, so if $T(c) = w$ and $L(w) = c$ then also $L(w') = c$. Note that PBT is a weaker protocol than $NCTD$, as we see in Figure 2 (left).

Somewhat in parallel with this evolution of consistency-based teaching, there have been some other views of machine teaching, from cases where experiments with humans are performed [11] to the use of machine teaching for explainable AI [20]. In this case, the teacher selects examples that maximize the explainee's probability of a correct inference. A teaching framework aimed at Bayesian learners is introduced in [22]. The framework is expressed as an optimisation problem over batch teaching examples that balance the future loss of the learner and the effort of the teacher. A new conceptualisation of *expected* teaching dimension using a learning and a sampling prior is presented in [10]. Shafto et al. [2014] present the idea that learning can be modelled as Bayesian inference, selecting a small subset of the data that will, with high probability, lead a learner model to the correct inference. A general framework for selecting examples to teach probabilistic learners is presented in [4]. Yang and Shafto [2017] use a Bayesian approach where teacher and learner interact and converge on the likelihood of the data given the model on the teacher's side and the posterior of the model given the data on the learner's side inspired by iterative teaching [13]. Overall, these papers present an interactive, non-batch setting, do not consider the notion of cheating or do not calculate teaching dimensions. In the traditional machine teaching setting we follow, we assume that the teacher works in a batch mode and sends a witness (e.g., a set of examples) once and for all. Even our use of the term 'changing mind' is metaphorical, as the examples do not come incrementally.

In this paper we study probabilistic teaching models, where the teacher and learner share a generalisation of the consistency graph (which is based on the binary function $\models$ on pairs of concepts and ground elements, which derives into a binary function $\models$ on pairs of concepts and witnesses) in the form of a joint probability distribution of concepts and witnesses $p : C \times W \to [0, 1]$. In this paper we consider only finite concept classes. Even if probabilities could be exploited by teacher and learner to extract confidence in the identification or set different thresholds for $p(c|w)$, in this paper we will require unique identification. This reduces to following p, and since $p(c|w) = p(c,w)/p(w)$ this means:

$$L(w) = \arg!\max_{c \in C} p(c|w) = \arg!\max_{c \in C} p(c,w) \tag{1}$$

where $\arg!\max$ only returns an element if it is unique, otherwise $L(w)$ is undefined (recall that L is a partial mapping). Now, for the teacher, following p means that if $T(c) = w$ we must have $p(c|w) > p(c'|w)$ for all concepts $c' \neq c$, as the teacher assumes that the learner simply follows the posterior and can identify one concept uniquely with it.

3. MT Models as Witness Joint Distributions

The extension of the consistency graph into a joint distribution assumes that if c and w are inconsistent then $p(c,w) = 0$, but if c and w are consistent, then $p(c,w) > 0$ could take any possible value, provided, of course, it is well-defined, i.e., $\sum_{c \in C, w \in W} p(c,w) = 1$. Basically, the extension converts possibility into probability. The learner then follows:

$$p(c|w) = \frac{p(c,w)}{p(w)} = \frac{p(w|c)p(c)}{p(w)}$$

Table 1 shows a summary of possible cases, depending on the constraints on $p(w|c)$ and $p(c)$ to build estimators for $p(c|w)$. By uniform prior we mean $\forall c, c' : p(c) = p(c')$. By uniform likelihood we mean $\forall c, w, w', c \models w, c \models w' : p(w|c) = p(w'|c)$. Note that if w and c are inconsistent then $p(w|c) = 0$ because $p(w,c) = 0$. We define the coverage of c as $W_c = \{w : c \models w\}$. If this set is finite, then $p(w|c) = \frac{1}{|W_c|}$ in the uniform likelihood case.

When both the likelihood and concept prior are considered uniform we have a few situations already. The general case is actually a new machine teaching model, when coverage sizes $|W_c|$ differ between concepts. Only if the coverage size $|W_c|$ is the same for all c, then we have the extreme case $\forall c, c', w, w', c \models w, c' \models w' : p(w|c) = p(w'|c')$. This happens in some well-studied situations, such as any class of Boolean concepts when using both positive and negative examples, since then for any concept c and any $w \subseteq 2^X$ there is a unique assignment of negative and positive to elements of w that will be consistent with c. This case really corresponds to the classical teaching (CT) dimension model of Goldman and Kearns [1995]: no preference exists between posteriors and the learner is undefined unless there is unique consistency. If the coverage size is not equal, then the likelihood is higher the smaller the coverage of the concept is, and this

Case	Existing and New Specific MT Models	$p(w\|c)$	$p(c)$	Results
Uniform Prior and Likelihood	CT [8] (concepts with same coverage), PBT [7] (concepts with smaller coverage prevail)	Uniform	Uniform	Monotone and GM-collusion-free
Uniform Likelihood	PBT [7] and Learning Prior [10] (if all concepts same coverage or extreme priors)	Uniform	Free	Monotone and GM-collusion-free
Uniform Prior	MLE Teaching	Free	Uniform	Lower bound on TD
Free	MAP Teaching	Free	Free	Lower bound on TD

Table 1: Four different new teaching models depending on constraints on likelihood or the concept prior as per Bayes' rule.

would be a specific case of the preference-based teaching (PBT) model of [7, 14], with smaller concepts (in coverage) having preference.

When only the likelihood is assumed uniform we are in a situation that follows the concept prior $p(c)$ and the coverage size of the concept when choosing among several consistent hypotheses. Again, if coverage sizes are equal, we are clearly in the PBT model again, with the priors leading directly to the preferences. However, if coverage sizes are not equal, we can still have an equivalent PBT model that follows the priors by choosing them in an extreme way such that the effect of the likelihood does not affect the choice from the posterior[1]. For instance, Occam's razor, which selects the simpler one of any two consistent hypotheses, could be represented in this way.

When only the concept prior is assumed uniform we are in a situation where

[1]We would have to choose a ratio in the priors so high to beat any effect of the likelihoods, i.e, $\forall c, c'$, if $p(c) > p(c')$ then $p(c)/p(c') > k$ such that k is greater than any likelihood involving these two concepts.

the “maximum likelihood estimation” (MLE) is used. Finally, in the general case where both the likelihood and concept prior can vary, we are in the most general case, “maximum a posteriori” (MAP) estimation.

For the two first rows in Table 1, but not the following two rows, we have that they are GM-collusion-free.

Proposition 1. *If likelihood is uniform then a learner L based on the posteriors is GM-collusion-free.*

Proof. In general, $p(c|w) = \frac{p(w|c)p(c)}{p(w)}$ so that $p(c|w) > p(c'|w) \Leftrightarrow p(w|c)p(c) > p(w|c')p(c')$. Thus if the likelihood is uniform then for any two witnesses w, w' both consistent with c and c', if we have $p(c|w) > p(c'|w)$ then also $p(c|w') > p(c'|w')$, which means that no change of mind can occur for a learner acting on the posteriors, i.e. we have GM-collusion-freeness. $\square$

The notion of GM-collusion-freeness is related to the intuitive principle that increasing consistent evidence should reinforce our beliefs. In a non-probabilistic setting, this is understood as not changing mind for any superset, but this is too extreme an interpretation. A more natural interpretation, which we call the monotonicity property, is that *the more examples a learner is given that are consistent with a concept, the more plausibility the learner should assign to that concept*, as it rules out other concepts. However, in a probabilistic setting, the probability for a concept c_1 can still increase (or not decrease) but another competing concept c_2 can increase its probability more, now beating the first. We can translate the monotonicity property to our probabilistic setting as follows. We say that p is monotone iff:

$$\forall c \, \forall w, w' \in W : c \models w \wedge w' \subseteq w \Rightarrow p(c|w') \leq p(c|w)$$

Note that the above property and the definition of GM-collusion-free are very similar. For the two first rows in Table 1 we show that monotonicity is preserved, but this does not hold for the last two rows.

Proposition 2. *If likelihood is uniform then p is monotone.*

Proof. Assume $c \models w$ and $w' \subseteq w$. By Bayes Rule and the uniform likelihoods:

$$\frac{p(c|w)}{p(c|w')} = \frac{p(w|c)p(c)/p(w)}{p(w'|c)p(c)/p(w')} = \frac{p(w')}{p(w)}$$

Also, since the likelihood is uniform, and w' is consistent with at least the same concepts as w, the marginal $p(w') = \sum_c p(w'|c)p(c) \geq \sum_c p(w|c)p(c) = p(w)$. So we have that $p(c|w) \geq p(c|w')$, which shows the monotonicity property. $\square$

Concept	$p(c)$	$p(c, \emptyset)$	$p(c, \{3\})$	$p(c, \{3, 11\})$	...
c_{even}	0.35	0.15	0	0	...
c_{odd}	0.35	0.15	**0.022**	0.004	...
c_{pow3}	0.1	0.08	0.013	0	...
c_{prime}	0.2	0.125	0.013	**0.005**	...

Table 2: Part of a witness joint probability with $|C| = 4$ over $W = 2^{\mathbb{N}}$ with decreasing probabilities for larger sets and larger probabilities for simpler concepts. As a result, we have a 'change of mind' between $\{3\}$ and $\{3, 11\}$, as the identified concept (in boldface) changes from c_{odd} to c_{prime}. This seems natural, as 11 is more specifically prime than odd (higher likelihood $p(w|c)$, not shown), but this is not GM-collusion-free.

To see why we need to go beyond the first two rows of Table 1, we show in Table 2 an example where changing mind is not necessarily cheating (even if not GM-collusion-free). This is a situation where neither $p(w|c)$ nor $p(c)$ are uniform, but a similar example can be found with $p(c)$ uniform.

We turn to proving results for the last two rows of Table 1, and show that the degree of freedom they allow is very high. We start by defining the minimum teaching dimensions achievable for these new paradigms when likelihoods can be freely chosen (MLE and MAP teaching).

Definition 1. *For C on ground elements X and $p : C \times 2^X$ a joint distribution and $c \in C$, let $JDTD^+(C, p, c)$ be the minimum cardinality of some $w \in 2^X$ such that $c = \arg!\max_{c \in C} p(c|w)$.*

Let $JDTD^+(C, p)$ be the maximum $JDTD^+(C, p, c)$ over all $c \in C$.

Let $JDTD^+(C)$ be the minimum $JDTD^+(C, p)$ over all p.

Analogously, we define $JDTD^{++}(C)$ for $p : C \times X^*$, i.e. multisets.

How powerful are these new paradigms? Does the use of any joint distribution p that is restricted only by $p(c, w) > 0 \Leftrightarrow c \models w$ allow us to reach the minimum possible teaching dimension in all situations? The answer is Yes. We first define a theoretical lower bound on teaching dimension.

Definition 2. *For C on ground elements X and family of witnesses W, and positive integer k, let $G^k(C)$ be the bipartite graph with a vertex for each concept $c \in C$ and a vertex for each $w \in W$ of at most $k > 0$ elements from X, and with an edge cw whenever $c \models w$.*

Define $LBTD^+(C)$ (for $W = 2^X$, i.e. sets) and $LBTD^{++}(C)$ (for $W = X^*$, i.e. multisets), as the minimum k such that $G^k(C)$ has a matching saturating C.

Note $LBTD^+ \leq LBTD^{++}$ as the former is defined over a subgraph of the latter. For any teacher mapping T where $c \models T(c)$ we have these variants of $LBTD(C)$ being a *Lower Bound on the Teaching Dimension* k achieved by T, as the edges $\{cT(c)\}_{c \in C}$ will form a matching saturating C in the graph $G^k(C)$.

Proposition 3. *$JDTD^+ = LBTD^+$ and $JDTD^{++} = LBTD^{++}$, even for distributions with uniform prior. For any concept class C on positive examples, with or without repetitions, there is a uniform prior joint distribution p such that a learner acting on posteriors achieves lowest possible teaching dimension.*

Proof. Assume $k = LBTD^+(C)$, or $k = LBTD^{++}(C)$, as per Definition 2 and consider the graph $G^k(C)$ with the set of k matching edges M saturating C. We construct p by assigning values to a joint distribution matrix $C \times W$, which we assume is an n by m matrix. We partition the nm values $p(c, w)$ into 3 classes: those where $c \not\models w$ which we set to $p(c, w) = 0$, those where $cw \in M$, and the remaining. To a concept c consistent with d witnesses we set $p(c, w) = \frac{2}{n(d+1)}$ for the unique w such that $cw \in M$, and $p(c, w) = \frac{1}{n(d+1)}$ for the remaining $d - 1$ witnesses consistent with c. Note that the marginals for each of the n concepts (rows) is $1/n$, so this is a uniform prior joint distribution. The teacher mapping follows the matching, with $T(c) = w$ for $cw \in M$. Given $T(c) = w$ the learner will follow the posteriors and since $p(c', w) < p(c, w)$ for all $c' \neq c$ the learner will correctly guess c. $\square$

We have seen that for the two new machine teaching paradigms MLE and MAP in Table 1, and any Boolean concept class, the theoretically lowest possible TD can be achieved by cherry picking the joint distribution. While it may be the case that these arbitrary distributions are actually the true information about the world that teacher and learner share, for an external observer this is impossible to tell. In order to clarify this, we now take a step back and define a class of joint distributions over witness sets where the distribution is constructed in a meaningful way.

4. The Witness Sampling Composition Model

The original notion of consistency is defined between concepts in C and ground elements in X. As we said at the beginning of section 2, if c is consistent with x_1 and x_2 then c must be consistent with $\{x_1, x_2\}$. Inconsistencies are also extended from X to W. So, does it make sense that p, the witness joint distribution, is not

Concept	$q(c)$	$q(c,3)$	$q(c,11)$	...
c_{even}	0.35	0	0	...
c_{odd}	0.35	0.087	0.0054	...
c_{pow3}	0.1	0.05	0	...
c_{prime}	0.2	0.05	0.0062	...

Table 3: Part of a joint distribution q with $|C| = 4$ on ground set $X = \mathbb{N}$. With regularisation terms $r(0) = 0.5$, $r(1) = 0.25$, $r(2) = 0.125, \ldots$ and using WSC without replacement this q gives the witness joint distribution p in Table 2.

an extension of the ground joint distribution q? For instance, if $q(c, x_1) < q(c', x_1)$ and $q(c, x_2) < q(c', x_2)$, does it make sense that $p(c, \{x_1, x_2\}) > p(c', \{x_1, x_2\})$?

In order to derive $p : C \times W \to [0, 1]$ from $q : C \times X \to [0, 1]$, we are going to assume the following: when two or more elements in X are composed in W their composition is performed as a sampling process. We call this assumption Witness Sampling Composition (WSC), and we define it as follows: WSC means that the construction of witnesses is modelled as a sampling procedure from X where the extraction of one element does not affect the relative probabilities of extracting the remaining elements (with or without replacement). We define the WSC construction of p from q recursively, with the base case given by $p(c, \lambda) = r(0) \cdot q(c)$ where λ represents the empty witness (no example has been sampled yet) and r is a regularisation term (with $\sum_n r(n) = 1$) we will explain later. The recursive step is defined as follows:

$$
\begin{aligned}
p(c, w) &= r(|w|) \cdot \sum_{x_i : w = w' \oplus x_i} \left[\frac{p(c, w')}{r(|w'|)} \cdot q(x_i | c) \right] \\
&= \frac{r(|w|)}{r(|w| - 1)} \cdot \sum_{x_i : w = w' \oplus x_i} \frac{p(c, w') \cdot q(c, x_i)}{\sum_{x \in X^{-w'}} q(c, x)}
\end{aligned} \tag{2}
$$

where $|w|$ represents the dimension of w (number of elements in w), and $w = w' \oplus x_i$ represents that witness w is composed of a smaller witness w' (of dimension $|w| - 1$) and $x_i \in X$. Finally, with $x \in X^{-w'}$ we denote any x that can be sampled from X after w' has been sampled (with replacement or not). Note the difference between the first and second line of the previous derivation is just a normalisation keeping the proportions (which is $q(c)$ when there is replacement).

Proposition 4. *Under WSC, the concept priors are preserved between q and p, i.e.: $\forall c \in C : p(c) = q(c)$*

Proof. We have $p(c) = \sum_{w \in W} q(c, w)$ by definition. We first prove, by induction on i, this Claim:

$$\sum_{w:|w|=i} p(c, w) = r(i)q(c)$$

The base case $i = 0$ of the Claim follows from the base case given right before Eq. 2 of the recursive definition of p from q by WSC, which says that $p(c, \lambda) = r(0)q(c)$.

For the induction step of the Claim we apply Eq. 2 to get

$$\sum_{w:|w|=n} p(c, w) =$$

$$= \sum_{w:|w|=n} r(n) \cdot \sum_{x_i : w = w' \oplus x_i} \left[\frac{p(c, w')}{r(n-1)} \cdot q(x_i | c, w') \right]$$

$$= r(n) \sum_{w':|w'|=n-1} \sum_{x_i : w = w' \oplus x_i} \left[\frac{p(c, w')}{r(n-1)} \cdot q(x_i | c, w') \right]$$

$$= r(n) \sum_{w':|w'|=n-1} \frac{p(c, w')}{r(n-1)} \sum_{x_i : w = w' \oplus x_i} [q(x_i | c, w')]$$

$$= r(n) \frac{1}{r(n-1)} \sum_{w':|w'|=n-1} p(c, w')$$

and applying the inductive assumption for $n - 1$, we get:

$$\sum_{w:|w|=n} p(c, w) = r(n) \frac{1}{r(n-1)} r(n-1)q(c) = r(n)q(c)$$

which completes the proof of the Claim.

Now, since $\sum_n r(n) = 1$, applying the Claim, we have:

$$p(c) = \sum_{w \in W} p(c, w) = \sum_n \sum_{w:|w|=n} p(c, w) =$$

$$= \sum_n r(n)q(c) = q(c)$$

and we are done with the proof of the proposition.

□

The meaning of the regularisation term r comes from the fact that $\forall c \forall n \geq 0$: $\sum_{w \in W:|w|=n} p(c,w) = r(n) \cdot q(c)$ but, as $q(c) = p(c) = \sum_{w \in W} p(c,w)$, then r is actually a regularisation probability $r : \mathbb{N} \to [0,1]$, where $r(n)$ represents how likely all the witnesses of dimension n are. In other words, if we are given the ground joint distribution q, expressing how likely any pair of concept and witness is, and we are also given how likely each dimension is, then we can derive the witness joint distribution. The choice of r does not affect the behaviour of the learner ($\arg!\max_c p(c|w)$), as when w is given, we have the same $r(|w|)$ for all concepts.

Concept	$p(c)$	$q(c,H)$	$q(c,T)$	$p(c,\{H\})$	$p(c,\{T\})$	$p(c,\{HT\})$	$p(c,\{HTT\})$
$coin_1$	0.25	0.2	0.05	**0.05**	0.0125	0.005	0.0005
$coin_2$	0.25	0.125	0.125	0.031	0.031	**0.0078**	0.0020
$coin_3$	0.25	0.10	0.15	0.025	0.037	0.007	**0.00225**
$coin_4$	0.25	0.05	0.2	0.0125	**0.05**	0.005	0.0020

Table 4: Coin example with $|C| = 4, |X| = 2$ and $W = X^*$. A ground joint probability distribution q and part of the WSC derived witness probability p of 4 biased coins with uniform $p(c)$. Coin 1: heavily biased Heads. Coin 2: Fair. Coin 3: mildly biased Tails. Coin 4: heavily biased Tails. p derived by WSC sampling with replacement and using $r(0) = 0.5, r(1) = 0.25, r(2) = 0.125, ...$ Note $TD = 3$.

We give two examples of ground distributions q and the resulting WSC derived witness distributions p. One for subsets of natural numbers and sampling without replacement, with q in Table 3 and p in Table 2, and one example with coins illustrating sampling with replacement in Table 4.

5. Teaching Dimension of the WSC Model

This section shows the remaining inequalities in Figure 1 with focus on the Sampling Teaching Dimension STD, where we allow full freedom in the choice of ground distribution q.

Definition 3. *For C over ground elements X and joint distribution $q : C \times X$, let $p_q : C \times 2^X$ be the joint distribution derived by WSC from q.*

Define $STD^+(C,q) = JDTD^+(C,p_q)$ and let $STD^+(C)$ be the minimum $STD^+(C,q)$ over all $q : C \times X$, with $STD^{++}(C)$ the analogous for multiset witnesses.

We start with the case of sets. Clearly we have $JDTD^+ \leq STD^+$ as the latter is a specialization of the former, but is it a strict restriction? We can answer this in the affirmative with a simple proof.

Proposition 5. *There is a concept class C where $JDTD^+(C) < STD^+(C)$.*

Proof. Consider $X = \{x_1, x_2\}$ and $C = \{c_1, c_2, c_3, c_4\}$ with $W = 2^X$. We have 4 concepts and 4 witnesses. With a free joint distribution we can simply do our 4×4 cells as $p(c_1, \emptyset) = 0.25 - \epsilon/4$, $p(c_2, \{x_1\}) = 0.25 - \epsilon/4$, $p(c_3, \{x_2\}) = 0.25 - \epsilon/4$, $p(c_4, \{x_1, x_2\}) = 0.25 - \epsilon/4$ and $p(c, w) = \epsilon/12$ for the other 12 combinations, with ϵ a sufficiently small number. $JDTD^+$ is hence 2.

Now, let us try to think of a possible ground joint distribution q, of dimension 2×4, to get the same teaching dimension when deriving p using WSC. Here, for each c we have that p is simply built from q by constructing a set of elements $w \subset X$ using sampling without replacement from X. By Eq. 2, we have for the empty set:

$$p(c, \emptyset) = r(0) \cdot q(c)$$

and for a set of size 1:

$$p(c, \{x\}) = r(1) \cdot q(c, x)$$

and a set of size 2:

$$p(c, \{x_1, x_2\}) = r(2) \cdot \left[\frac{q(c, x_1)q(c, x_2)}{\sum_{x \neq x_1} q(c, x)} + \frac{q(c, x_2)q(c, x_1)}{\sum_{x \neq x_2} q(c, x)} \right] =$$

$$= r(2) \cdot \left[\frac{q(c, x_1)q(c, x_2)}{q(c, x_2)} + \frac{q(c, x_2)q(c, x_1)}{q(c, x_1)} \right] =$$

$$= r(2) \cdot [q(c, x_1) + q(c, x_2)] = r(2) \cdot q(c)$$

And we see that the concept choice for both $w = \emptyset$ and $w = \{x1, x2\}$, since $r(0)$ and $r(2)$ are constants, is dominated by $q(c)$. Thus, we will have $\arg!\max_c(p(c, \emptyset)) = \arg!\max_c(p(c, \{x1, x2\}))$. Thus, 2 of the 4 witnesses cannot both be used to distinguish between concepts, and so we must have $STD^+(C) > 2$. □

Our next result shows that STD^+, even when restricted to uniform prior distributions, is as powerful as any non-clashing teaching model. To prove this proposition we construct a ground joint distribution that cherry-picks the non-clashing teacher function, by assigning small values to $q(c, x)$ if $c \models x$ but $x \notin T(c)$. Note however that these values are not exponentially small, as they satisfy $q(c, x) > 1/|C|^3$.

Proposition 6. $STD^+ \leq NCTD^+$. *For any* C *we have a uniform prior distribution* $q : C \times X$ *such that* $p : C \times 2^X$ *derived by WSC from* q *shows that* $STD^+(C) \leq NCTD^+(C)$.

Proof. Consider some C on ground set X and set of witness objects W (positive examples only). Assume a teacher mapping $T : C \to 2^X$ and for all $c \neq c'$ either $c \not\models T(c')$ or $c' \not\models T(c)$, i.e. non-clashing/GM-collusion-free.

We assign values to a joint distribution matrix $q : C \times X$. Assume $|C| = n$ and $|X| = m$. We will construct an assignment so that $q(c) = 1/n$ for all $c \in C$, thus uniform priors, and so that a learner following the posteriors of the WSC without replacement derived distribution $p : C \times 2^X$ will correctly guess c when given the witness set defined by $T(c)$, to prove the proposition.

Let $\epsilon > 0$ be a small value to be decided later. Consider some $c \in C$ and assume that $|T(c)| = k$ and that c consistent with a further d ground elements. Assuming $d > 0$ we assign: $q(c, x) = 0$ for $c \not\models x$, and $q(c, x) = \epsilon/d$ for the d ground elements consistent with c but not in $T(c)$. Note these values sum to ϵ, so we have $1/n - \epsilon$ left to assign for this c and we do this by setting $q(c, x) = (1/n - \epsilon)/k = \frac{1-\epsilon n}{kn}$ for each ground element $x \in T(c)$. If $d = 0$ we assign: $q(c, x) = 0$ for $c \not\models x$, and $q(c, x) = 1/(kn)$ for each $x \in T(c)$.

To prove the proposition we show that the joint distribution $p : C \times 2^X$ derived by WSC from this q will have the following main property: "for any two concepts $c \neq c'$ we have $p(c, T(c)) > p(c', T(c))$". Let $|T(c)| = k$. We have 2 cases:

(1): $c' \not\models T(c)$. Then $p(c', T(c)) = 0$ and we are done.

(2): not (1) so since T is non-clashing we have $c \not\models T(c')$. Assume $|T(c') - T(c)| = t$ and $|T(c) - T(c')| = s$. We have $t \geq 1$ and $s \geq 0$. We have k ground elements in $T(c)$, and $q(c, x) = \frac{1-\epsilon n}{kn}$ for all, while for s of them $q(c', x)$ will be at most ϵ/s, and since $|T(c')| \geq k + 1 - s$ then for the remaining $k - s$ of them we have $q(c', x) \leq q(c, x)\frac{k}{k+1-s}$. We thus have $\Sigma_{x \in T(c)} q(c, x) = \frac{1-\epsilon n}{n}$ and $\Sigma_{x \in T(c)} q(c', x) = \frac{1-\epsilon n}{n} \times \frac{k-s}{k-s+1} + \epsilon$. Since we can choose $\epsilon > 0$ as low as we want, we now have a situation where the k values for $q(c, x)$ are all equal and both their sum and their product, respectively, is larger than the sum and the product, respectively, of the k values $q(c', x)$. Since q has uniform priors and hence by Proposition 4 also p has uniform priors, we therefore must have that the WSC computation of probabilities when sampling without replacement, gives that $p(c, w) > p(c', w)$. Note that choosing $\epsilon = 1/|C|^2$ suffices, as we then will have the first sum (values for c) being $(n - 1)/n$ and the second sum (values for c') being $((n - 1)^2 + 1)/n^2$. □

Now we turn to the set of witness objects being multisets over X. Firstly, as in the set case, we clearly have $JDTD^{++} \leq STD^{++}$ and there is an easy proof showing that STD^{++} is not as free as $JDTD^{++}$.

Proposition 7. *There is a concept class C where $JDTD^{++}(C) < STD^{++}(C)$.*

Proof. Consider any $|C| = 5$ and $|X| = 2$ with $T : C \to X^*$ using 2 witnesses of size 1 and 3 witnesses of size 2, as can be done with the completely free choice of p allowed by JDTD to give $JDTD^{++}(C) = 2$. As WSC is not able to achieve both $p(c, \{x, x\}) > p(c', \{x, x\}$ and $p(c', \{x\}) > p(c, \{x\}$ we must have $STD^{++}(C) > 2$. □

This proof actually shows that STD avoids a very unnatural situation which looks like cheating, e.g. tossing a coin where Tail is more likely than Head but two Heads more likely than two Tails, and this is forbidden by STD. Finally, we show a somewhat surprising result comparing teaching dimensions using witness objects that are multisets versus sets, showing that STD^{++} restricted to uniform prior distributions will achieve the theoretical lower bound $LBTD^{+}$. To prove this proposition we construct a ground joint distribution that cherry-picks any given matching in $G^k(C)$ as per Definition 2, by assigning small values to $q(c, x)$ if $c \models x$ but $x \notin T(c)$. These values satisfy $q(c, x) > 1/(|C||X|)^2$.

Proposition 8. $STD^{++} \leq LBTD^{+} = JDTD^{+}$. *For any concept class C over X and set of witness objects $W = 2^X$, there exists a uniform prior joint distribution $q : C \times X$ such that $p_q : C \times X^*$ derived by WSC with replacement from q has teaching dimension $STD^{++}(C) \leq LBTD^{+}(C) = JDTD^{+}(C)$.*

Proof. The latter equality follows from Proposition 3. Assume $k = LBTD^{+}(C)$ as by Definition 2 and consider the graph $G^k(C)$ with the set of matching edges $M = \{cw\}_{c \in C}$. We assign values to a joint distribution matrix $q : C \times X$. Assume $|C| = n$ and $|X| = m$. We will construct an assignment so that $q(c) = 1/n$ for all $c \in C$, thus uniform prior, and so that a learner following the posteriors of the WSC with replacement derived distribution $p : C \times X^*$ will correctly guess c when given the witness set defined by $T(c) = w : cw \in M$, to prove the proposition.

Let $\delta = 1/(nm^2)$. Consider some $c \in C$ and assume that $|T(c)| = k$ and that c consistent with a further d ground elements. Assuming $d > 0$ we assign: $q(c, x) = 0$ for $c \not\models x$, and $q(c, x) = \delta/d = 1/(nm^2 d)$ for the d ground elements consistent with c but not in $T(c)$. Note these values sum to δ, so we have $1/n - \delta$ left to assign for this c and we do this by setting $q(c, x) = (1/n - \delta)/k = (m^2 - 1)/(knm^2)$ for each ground element $x \in T(c)$. If $d = 0$ we assign: $q(c, x) = 0$ for $c \not\models x$, and $q(c, x) = 1/(kn)$ for each $x \in T(c)$.

To prove the proposition we show that the joint distribution $p : C \times X^*$ derived by WSC- witness sampling composition - with replacement - from this q will have the following main property: "for any two concepts $c \neq c'$ we have $p(c, T(c)) > p(c', T(c))$". Let $T(c) = \{x_1, x_2, ..., x_k\}$. Since p is defined by doing sampling with replacement, and since q has uniform priors and hence by Proposition 4 also p has uniform priors, to prove $p(c, T(c)) > p(c', T(c))$ it suffices to show that $\Pi_{i=1}^{k} q(c, x_i) > \Pi_{i=1}^{k} q(c', x_i)$. We have 3 cases.

(1): if there exists $x \in T(c)$ such that $c' \not\models x$ then $q(c', x) = 0$ so that $p(c', T(c)) = 0$ and we are done.

(2): not (1) but we have $T(c) \subset T(c')$ and thus $|T(c)| < |T(c')|$. We settle this case by showing that for every $x \in T(c)$ we have $q(c', x) < q(c, x)$. We observe that $q(c', x) \leq 1/((k+1)n)$ and $q(c, x) \geq (m^2-1)/(knm^2)$ for any $x \in T(c)$, so the inequality that needs to be shown, after rearranging, is that $m^2/(m^2-1) < (k+1)/k$ and this holds since $k < m$.

(3): not (1) or (2), so we have some $x' \in T(c) - T(c')$ with $c' \models x'$. Let $|T(c) - T(c')| = s$. To show $p(c, T(c)) > p(c', T(c))$ the hardest case is when all the $q(c', x)$ values for $x \in T(c)$ are as high as possible. This will occur when $s = 1$ and $T(c') \subset T(c)$ so $|T(c')| = k-1$, and we have $q(c', x) = (m^2-1)/((k-1)nm^2)$ for $x \in T(c') \cap T(c)$ and $q(c', x_i) = 1/(nm^2)$ for the unique $x_i \in T(c) - T(c')$, while $q(c, x) = (m^2-1)/(knm^2)$ for all $x \in T(c)$. Thus we need to show

$$\frac{1}{nm^2} \times \frac{(m^2-1)^{k-1}}{((k-1)nm^2)^{k-1}} < \frac{(m^2-1)^k}{(knm^2)^k}$$

After rearranging this resolves to showing $k \times (k/(k-1))^{k-1} < m^2 - 1$. Note we can assume that $m > k$, as $m = k$ would imply that both c and c' consistent with every ground element and hence $c = c'$. Thus we need to show $k \times (k/(k-1))^{k-1} < (k+1)^2 - 1$, which holds since with $k' = k-1$ we have $(k/(k-1))^{k-1} = ((k'+1)/k')^{k'} = (1+1/k')^{k'} < e < 2.72$ and thus we need to show $2.72k < (k+1)^2 - 1$ which holds for any $k \geq 1$.

Thus for all 3 cases, we have shown for any pair of concepts $c \neq c'$ that $p(c, T(c)) > p(c', T(c))$, and we are done with the proof of the proposition. □

6. Computational Complexity

Let us now consider the computational complexity of computing the newly introduced teaching dimensions. Note that some of the other teaching dimensions are computationally hard. For instance, Kirkpatrick et al. [12] show that already deciding whether $NCTD^+(C) = 1$ is NP-hard. We consider first the case of computing $LBTD^+$, or equivalently $JDTD^+$ by Proposition 3, i.e. the smallest teaching dimension achievable when teaching sets must obey consistency. We emphasize that we assume non-redundant concept classes, i.e. for a concept class C over X we have no two concepts consistent with the exact same subset of X, and moreover for each element of X there is at least one concept consistent with it and another not consistent with it. By Proposition 3 we have that $JDTD^+(C) = LBTD^+(C)$. By Definition 2 deciding the value of the latter resolves to finding the smallest value k such that $G^k(C)$ has a matching saturating C. Recall that the bipartite graph $G^k(C)$ has one color class being C and the other color class being the set of witnesses on at most k positively labelled examples, and with edges denoting consistency.

Lemma 9. *The graph $G^{|X|}(C)$ will always have a matching saturating C.*

Proof. Note first that all concepts in C are distinct, meaning that no two of them are consistent with exactly the same set of positively labelled examples from X. As the graph $G^{|X|}(C)$ contains all witnesses of size up to $|X|$ no two vertices from C has the same set of neighbors. Now consider any $B \subseteq C$ and let $N(B)$ be the set of witnesses consistent with at least one concept from B. By induction on $|B|$ it follows that $|N(B)| \geq |B|$, with the base case $|B| = 1$ clear (the empty witness set is also a vertex), and for the induction take a vertex $c \in B$ and note that c has a neighbor not shared with any other vertex in B. Now, by matching theory we know that if for all $B \subseteq C$ we have $|N(B)| \geq |B|$, then a matching saturating C exists. $\square$

Theorem 10. *Given a finite concept class C over a finite domain X, we can compute $LBTD^+(C) = JDTD^+(C)$ in time $O(\log k(|C| + |X|^k)^{5/2})$, where $LBTD^+(C) = k$.*

Proof. As discussed above we need to find the smallest value of k for which $G^k(C)$ has a matching saturating C. For a fixed value of k, to check if $G^k(C)$ has a matching saturating C we first construct the graph $G^k(C)$ in time $O(|C||X|^k)$ and then run the Hopcroft-Karp $O(|V|^{5/2})$ algorithm to check if the maximum matching has size $|C|$. We search for the smallest k by a variation of binary search called exponential search: first doubling until finding 2^i, the smallest power of two that has a positive answer, and then doing binary search between 2^{i-1} (which had a negative answer) and 2^i, for a total of $O(\log LBTD^+(C))$ trials. $\square$

Note that this algorithm shows that when parameterized by $JDTD^+(C)$ this problem is in the complexity class XP. Moreover, we have the following somewhat surprising result for concept classes that are big relative to the domain set.

Corollary 11. *If $|C|$ is lower-bounded by an exponential in $|X|$ then we can compute $LBTD^+(C)$ in polynomial time.*

We leave as an open problem the question if we can compute $LBTD^+(C)$ in polynomial time also when $|C|$ is not lower-bounded by an exponential in $|X|$. Note that defining $s(C)$ to be the smallest value k such that $\sum_{i=0}^{i=k} \binom{|X|}{i} \geq |C|$ then it is clear that we must have $LBTD^+(C) \geq s(C)$, simply since there must be enough witnesses to cover C. However, there are concept classes where we need more, e.g. take $X = \{1, 2, 3, 4\}$ and $C = \{\emptyset, 1, 2, 3, 4, 12, 14, 24, 124\}$. Note $|C| = 9$, $|D| = 4$ and thus $s(C) = 2$. In $G^2(C)$ we have a vertex for each of the 9 $c \in C$ and 11 vertices corresponding to all subsets of X of size at most 2, but 3 of these latter vertices, corresponding to $\{13, 23, 34\}$, have no neighbors in C and thus no matching saturating C can exist, so in this case we have $LBTD^+(C) > s(C)$.

Let us turn to $STD(C, q)$ and consider the complexity of computing the posteriors of $p_q : C \times W \to [0, 1]$ derived by Witness Sampling Composition from a given $q : C \times X \to [0, 1]$. Finding the unique concept maximizing the posterior $p_q(c|w)$ for a witness set w, to output $L(w)$, if it exists, is straightforward once we have computed p_q, following the argmax in Equation 1. Computing p_q from q for witnesses w that are multisets, i.e. using WSC with replacement, is a simple multiplication of the probabilities of each individual element of w. However, for WSC without replacement this computation is not linear in the cardinality of w, as it involves going over all permutations of the elements in w. To avoid the factor $|w|!$ that would thus arise, we can instead use dynamic programming for a factor $2^{|w|}$, see e.g. Chapter 2.3 of [16], but it seems the exponential runtime dependency on $|w|$ is unavoidable. If $|w|$ is not small, contrary to the intention of teaching, we could instead use a heuristic by assuming the elements are sampled independently, giving a close approximation if the distribution is nearly uniform, but the right approach here should be discovered experimentally.

7. Discussion

The paper of Kirkpatrick et al [12] was remarkable, clarifying the limits of the classical view of cheating as GM-collusion, by reinterpreting it as non-clash machine teaching. This seemed the culmination of over twenty years work of finding more and more powerful GM-collusion-free models, i.e., the most powerful non-cheating teaching models. However, we have challenged this very notion of cheating, under a natural probabilistic view. The notion of full-proof identification is replaced by the inductive notion of a guess, and the learners can 'change their mind' as posteriors are affected by changing likelihoods. We have also shown that in some particular cases (e.g., dealing with the empty set), the GM-collusion-free paradigm can allow some unnatural assignments.

The use of a witness joint probability p chosen freely has been shown powerful enough to equal the best achievable teaching dimension, $LBTD$, and in some cases, it can also do some unnatural assignments. Instead, we have proposed to derive p from the ground joint distribution q using WSC, in the same way as the consistency graph for witnesses derives from the consistency graph for ground elements, depending on how witnesses are composed. The new teaching models deriving from WSC are shown to be slightly less powerful than $LBTD$, but avoid the unnatural situations.

As a result, whether a teacher and learner do any cheating depends on whether the ground joint distribution corresponds to the joint beliefs about the world and how this distribution is extended to the witness joint distribution. If a teacher and learner know that a coin has 73% bias for heads, using this information for identifying the coin is not cheating. An epistemic form of cheating would be if both agreed that this coin had a different bias from the actual, one more convenient for the identification. And a logical, and more obvious form of cheating that is not tolerated by WSC would be to consider that the probability of seeing two heads is not 0.73^2 but some other probability more convenient for the identification.

Note that in the traditional conception of teaching dimension and collusion, the epistemic form of cheating is ignored altogether. For instance, any preference on the concepts that minimizes the teaching dimension would be GM-collusion-free, even if it does not correspond with the world. For instance, if in order to teach concepts over the natural numbers it is more advantageous that the concept of prime is preferred over the concept of odd, so be it, even if in the world where teacher and learner belong the concept of odd is usually more common. Of course, this situation also happens with the use of q, which still gives more freedom. This is why we emphasize that it is more natural to consider that q is given by the common knowledge that teacher and learner have about the world, and then the teaching dimension logically derives from this q, using the WSC procedure. The fact that the teaching dimension is determined by q and WSC does not mean that the procedure is necessary efficient for the teacher. This now becomes a more interesting question, namely asking about the computational complexity of computing the posteriors of $p : C \times W \rightarrow [0, 1]$ derived by WSC from a given factual q.

In this paper we have considered that teacher and learner share this joint probability q about the world, and the teaching process does not modify this q or the p that derives from it. But there are interactive situations that have already been studied in paradigms such as Bayesian teaching. These could consider cases such as the teacher and learner having different q_T and q_L, and an interactive process of teaching allowing them to exchange information about unresolved, incorrect or correct identifications, so that q_T and q_L can eventually converge. The use of entropy and other probability divergence metrics is not directly related to the teaching dimension, but these tools may be more useful to determine the uncertainty of the identification.

Overall, the general and refreshing probabilistic perspective synthesized in Table 1, and the connections with the classical teaching models, suggest that our paper could help bridge two very different conceptions of machine teaching in AI. There is the classical notion of identification, associated with theoretical results about the teaching dimension, and a more modern view of machine teaching as a probabilistic (or Bayesian) process. This should lead to future work connecting our schema to areas such as MDL/MML [18] inference, or when teacher and learner can adapt their probabilities, as mentioned above. All this can be explored with a more natural paradigm of non-cheating teaching that allows changes of mind.

Acknowledgments

We would like to thank Pekka Parviainen for providing insights that led to the witness sampling composition. We thank the anonymous reviewers for their comments and interaction during the discussion process. This work has been partially supported by the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, also by the EU (FEDER) and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and “ERDF A way of making Europe” Generalitat Valenciana under grant PROMETEO/2019/098, EU’s Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), and INNEST/2021/317 (Project cofunded by the European Union with the “Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER) de la Comunitat Valenciana 2014-2020”).

References

[1] Frank J Balbach. Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, 397(1-3):94–113, 2008.

[2] Thorsten Doliwa, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *International Conference on Algorithmic Learning Theory*, pages 209–223. Springer, 2010.

[3] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *The Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

[4] Baxter S Eaves-Jr. and Patrick Shafto. Toward a general, scaleable framework for bayesian teaching with applications to topic models. *CoRR*, 2016.

[5] Cèsar Ferri, José Hernández-Orallo, and Jan Arne Telle. Non-cheating teaching revisited: A new probabilistic machine teaching model. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2973–2979. ijcai.org, 2022.

[6] Ziyuan Gao, David Kirkpatrick, Christoph Ries, Hans Simon, and Sandra Zilles. Preference-based teaching of unions of geometric objects. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 185–207. PMLR, 15–17 Oct 2017.

[7] Ziyuan Gao, Christoph Ries, Hans Ulrich Simon, and Sandra Zilles. Preference-based teaching. *Journal of Machine Learning Research*, 18:31:1–31:32, 2017.

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

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

[10] Jose Hernández-Orallo and Jan Arne Telle. Finite and confident teaching in expectation: Sampling from infinite concept classes. In *ECAI 2020*, pages 1182–1189. IOS Press, 2020.

[11] Faisal Khan, Xiaojin Zhu, and Bilge Mutlu. How do humans teach: on curriculum learning and teaching dimension. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1449–1457, 2011.

[12] David Kirkpatrick, Hans U Simon, and Sandra Zilles. Optimal collusion-free teaching. In *Algorithmic Learning Theory*, pages 506–528. PMLR, 2019.

[13] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. Iterative machine teaching. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

[14] Farnam Mansouri, Yuxin Chen, Ara Vartanian, Jerry Zhu, and Adish Singla. Preference-based batch and sequential teaching: Towards a unified view of models. *Advances in Neural Information Processing Systems*, 32:9199–9209, 2019.

[15] Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Compressing and teaching for low VC-dimension. In *Symposium on Foundations of Computer Science*, pages 40–51, 2015.

[16] Pekka Parviainen. *Algorithms for exact structure discovery in Bayesian networks*. PhD thesis, 2012.

[17] Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55 – 89, 2014.

[18] C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.

[19] Scott Cheng-Hsin Yang and Patrick Shafto. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 workshop on Teaching Machines, Robots, and Humans*, pages 127–137, 2017.

[20] Scott Cheng-Hsin Yang, Wai Keen Vong, Ravi B Sojitra, Tomas Folke, and Patrick Shafto. Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, 11(1):1–17, 2021.

[21] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

[22] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *Neural Information Processing Systems 26*, pages 1905–1913. Curran, 2013.

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