Journal of Machine Learning Research 25 (2024) 1-34 Submitted 8/23; Revised 3/24; Published 3/24

MAP- and MLE-Based Teaching[*]

Hans Ulrich Simon HSIMON@MPI-INF.MPG.DE
Max-Planck Institute for Informatics, Germany
and Ruhr-University Bochum, Department of Mathematics, Germany

Jan Arne Telle TELLE@II.UIB.NO
Department of Informatics, University of Bergen, Norway

Editor: Aryeh Kontorovich

Abstract

Imagine a learner L who tries to infer a hidden concept from a collection of observations. Building on the work of Ferri et al. (2022), we assume the learner to be parameterized by priors $P(c)$ and by c-conditional likelihoods $P(z|c)$ where c ranges over all concepts in a given class C and z ranges over all observations in an observation set Z. L is called a *MAP-learner* (resp. an *MLE-learner*) if it thinks of a collection S of observations as a random sample and returns the concept with the maximum a-posteriori probability (resp. the concept which maximizes the c-conditional likelihood of S). Depending on whether L assumes that S is obtained from ordered or unordered sampling resp. from sampling with or without replacement, we can distinguish four different sampling modes. Given a target concept $c^* \in C$, a teacher for a MAP-learner L aims at finding a smallest collection of observations that causes L to return c^*. This approach leads in a natural manner to various notions of a MAP- or MLE-teaching dimension of a concept class C. Our main results are as follows. First, we show that this teaching model has some desirable monotonicity properties. Second we clarify how the four sampling modes are related to each other. As for the (important!) special case, where concepts are subsets of a domain and observations are 0,1-labeled examples, we obtain some additional results. First of all, we characterize the MAP- and MLE-teaching dimension associated with an optimally parameterized MAP-learner graph-theoretically. From this central result, some other ones are easy to derive. It is shown, for instance, that the MLE-teaching dimension is either equal to the MAP-teaching dimension or exceeds the latter by 1. It is shown furthermore that these dimensions can be bounded from above by the so-called antichain number, the VC-dimension and related combinatorial parameters. Moreover they can be computed in polynomial time.

1. Introduction

In formal models of *machine learning* we have a concept class C of possible concepts/hypotheses, an unknown target concept $c^* \in C$ and training data given by correctly labeled random examples. In formal models of *machine teaching* a collection $T(c^*)$ of labeled examples is instead carefully chosen by a teacher T in a way that the learner can reconstruct the target concept c^* from $T(c^*)$. In recent years, the field of machine teaching has seen various applications in fields like explainable AI as in Håvardstun et al. (2023), trustworthy AI as in Zhu et al. (2018) and pedagogy as in Shafto et al. (2014).

[*]. This research was supported by the Norwegian Research Council, project “Machine Teaching for XAI”.

Various models of machine teaching have been proposed, e.g. the classical teaching model in Shinohara and Miyano (1991) and in Goldman and Kearns (1995), the optimal teacher model in Balbach (2008), the recursive teaching in Zilles et al. (2011), the preference-based teaching in Gao et al. (2017), or the no-clash teaching in Kirkpatrick et al. (2019) and in Fallat et al. (2022). These models differ mainly in the restrictions that they impose on the learner and the teacher in order to avoid unfair collusion or cheating. The common goal is to keep the size of the largest teaching set, $\max_{c \in C} |T(c)|$, as small as possible.

There are also other variants using probabilities, from Muggleton (1996) where examples are sampled based on likelihoods for a target concept, to Shafto et al. (2014) who calls this pedagogical sampling and leads into the Bayesian Teaching of Eaves and Shafto (2016) and of Yang and Shafto (2017), to the Bayesian learners of Zhu (2013) with a proper teacher selecting examples.

In this paper we continue this line of research and consider the probabilistic model that had been described in the abstract. This model is inspired by and is an extension of the model that was introduced in Ferri et al. (2022). As already observed in Ferri et al. (2022), the condition for collusion-avoidance from Goldman and Mathias (1996) may here be violated, i.e., the learner may first reconstruct a concept c_1 from some given observations but, after having received additional observations, switch to another concept c_2 even if the new observations have given additional support to c_1. Like the authors of Ferri et al. (2022), we would like to argue that this should not be considered as collusion or cheating as long as the parameters assigned to the learner reflect some factual information about the world. In our paper the parameters can be freely chosen, and thus in order to disallow collusion one would have to impose more restrictions on the model, for instance a notion of natural parameterization. However, that is not our focus in this paper, which is rather on finding lower bounds on the teaching dimension that are even valid in the most liberal model of MAP-teaching.

As already outlined in the abstract, we will distinguish between four distinct sampling modes: ordered sampling with replacement ((O,R)-mode), unordered sampling with replacement ($(\overline{O},R)$-mode), ordered sampling without replacement ($(O,\overline{R})$-mode) and unordered sampling without replacement ($(\overline{O},\overline{R})$-mode). The smallest number d such that every $c^* \in C$ can be taught to a given MAP-learner L by a collection of at most d observations is denoted by $\text{MAP-TD}_L^{\alpha,\beta}(C)$ where $(\alpha,\beta) \in \{O,\overline{O}\} \times \{R,\overline{R}\}$ indicates the underlying sampling mode. Then $\text{MAP-TD}^{\alpha,\beta}(C) = \min_L \text{MAP-TD}_L^{\alpha,\beta}(C)$ is the corresponding parameter with an optimally parameterized learner L. The analogous notation is used for MLE-learners. Our main results are as follows:

1. The MAP-teaching model has two desirable and quite intuitive monotonicity properties. Loosely speaking, adding new observations (making Z larger) leads to smaller MAP-TD while adding new concepts (making C larger) leads to larger MAP-TD. See Section 3.2 for details.

2. The sampling modes (O,R) and $(\overline{O},R)$ are equivalent. The sampling modes $(\overline{O},R)$, $(O,\overline{R})$ and $(\overline{O},\overline{R})$ are pairwise incomparable (i.e., which one leads to smaller values of $\text{MAP-TD}_L(C)$ depends on the choice of C and L). Note that incomparability of the modes (α,β) and (α',β') does not rule out the possibility that $\text{MAP-TD}^{\alpha,\beta}(C) \le \text{MAP-TD}^{\alpha',\beta'}(C)$ for each concept class C. See Section 3.3 for details.

3. As for the (important!) special case, where concepts are subsets of a domain and observations are 0,1-labeled examples, we obtain some additional results, the first of which is the central one:

   (a) For a (properly defined) bipartite graph $G(C)^{\alpha,\beta}$ associated with C and $(\alpha, \beta) \neq (O, R)$, one gets[1]

   $$\text{MAP-TD}^{\alpha,\beta}(C) = \text{SMN}(G(C)^{\alpha,\beta}) \ . \tag{1}$$

   If we replace $G(C)^{\alpha,\beta}$ by a slightly modified graph, we obtain the corresponding result for MLE-TD at the place of MAP-TD.[2] Fig. 1 visualizes this result. See Sections 4 and 5.1 for details.

   (b) The MLE-teaching dimension is either equal to the MAP-teaching dimension or exceeds the latter by 1. See Section 5.2 for details.

   (c) The MAP- and the MLE-teaching dimension can be bounded from above by the so-called antichain number, the VC-dimension and related combinatorial parameters. See Section 5.3 for details.

   (d) Moreover the MAP- and the MLE-teaching dimension can be computed in polynomial time from a natural encoding of the underlying concept class. See Section 5.4 for details.

2. Definitions and Notations

We first fix some general notation. Afterwards, in Sections 2.1, 2.2, and 2.3, the MAP- and MLE-based teaching model is introduced, step-by-step.

Mappings. The restriction of a mapping $f : A \to B$ to a subset $A' \subseteq A$ will be denoted by $f_{\downarrow A'}$. Suppose that B is a set that is equipped with a size function which associates a size $|b|$ with each $b \in B$. Then the *order of a mapping* $f : A \to B$ is defined as the size of the largest element in the image of f, i.e., the order of f equals $\max_{a \in A} |f(a)|$.

Graphs and Matchings. For a graph $G = (V, E)$ and a set $U \subseteq V$, we denote by $\Gamma(U)$ the set of vertices which are adjacent to at least one vertex in U. If $G = (V_1, V_2, E)$ is the bipartite graph with vertex sets V_1 and V_2 and with edge set $E \subseteq V_1 \times V_2$, then $U \subseteq V_1$ implies (of course) that $\Gamma(U) \subseteq V_2$. A matching M in a bipartite graph $G = (V_1, V_2, E)$ can be viewed as a (partially defined and injective) function $M : V_1 \to V_2$ with the property that $(v, M(v)) \in E$ for each v having an M-partner. If V_1 is *saturated by* M, i.e., every vertex in V_1 has an M-partner, then this function is fully defined.

VC-Dimension, Vapnik and Chervonenkis (1971). Let C be a family of subsets of some ground set X. For $c \in C$ and $x \in X$, we also write $c(x) = 1$ if $x \in c$ and $c(x) = 0$ if $x \notin c$. We say that $S \subseteq X$ is *shattered by* C if, for every $b : S \to \{0, 1\}$, there is some $c \in C$ that coincides with b on S. The *VC-dimension of* C is defined as ∞ if there exist arbitrarily large shattered sets, and it is defined as the size of a largest shattered set otherwise.

[1] SMN(G) denotes the saturating matching number of a bipartite graph G (formally defined in Section 4)

[2] Some bounds on MLE-TD numbers in terms of SMN numbers are already found in Ferri et al. (2022), but no results that hold with equality (as in (1)) are proven there.

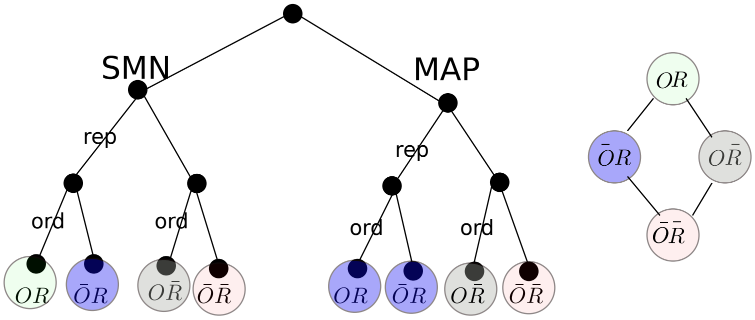

Figure 1: For any binary concept class $C \subseteq 2^X$ and $0,1$-labeled examples as observations, the tree visualizes the identities in (1). Using the same color for the two leftmost leaves in the MAP-subtree is justified by the equivalence of the modes (O, R) and $(\overline{O}, R)$ as shown in Corollary 19. A parameter represented by a leaf in the MAP-subtree has the same value as the parameter represented by a leaf of the same color in the SMN-subtree, as shown in Theorems 25, 27, 28 and Corollary 26. The parameters represented in the SMN-subtree are ordered as indicated by the rightmost diagram, with lowest value on top and highest value at bottom. We will see later, in Theorem 20, that parameters represented in different colors can generally have different values.

2.1 Concept Classes

Let C be a finite set of size at least 2, let Z be another non-empty finite set and let $\models$ be a relation on $C \times Z$. We refer to C as a *concept class* and to Z as a set of *observations*. If $c \models z$, then we say that the concept c is *consistent with the observation* z. We say that c is *consistent with a set (resp. multiset) A of observations*, which is written as $c \models A$, if c is consistent with every $z \in A$. The notation $c \models \mathbf{z}$ with $\mathbf{z} = (z_1, \ldots, z_n) \in Z^n$ is understood analogously. For each $c \in C$, we define

$$Z_c = \{z \in Z : c \models z\} \ .$$

Example 1 (Positive Examples as Observations) *Let $Z = X$ be a set of* examples *and let C be a family of subsets of X. Let the consistency relation be given by*

$$\forall c \in C, x \in X : c \models x \Leftrightarrow x \in c \ .$$

Note that $Z_c = c$ in this setting, i.e., concepts are identified with the sets of observations they are consistent with.

Example 2 (Labeled Examples as Observations) *Let $Z = X \times \{0, 1\}$ be a set of* labeled examples *and let C be a family of subsets of X. Let the consistency relation be given by*

$$\forall c \in C, (x, b) \in Z : c \models (x, b) \Leftrightarrow (b = 1 \wedge x \in c) \vee (b = 0 \wedge x \notin c) \ . \tag{2}$$

Note that $Z_c = \{(x, 1) : x \in c\} \cup \{(x, 0) : x \notin c\}$ in this setting. It follows that $|Z_c| = |X|$ for all $c \in C$.

We will occasionally identify a set $c \subseteq X$ with the corresponding $0,1$-valued function so that $c(x) = 1$ for $x \in c$ and $c(x) = 0$ for $x \in X \setminus c$. The equivalence in (2) can then be written in the form $c \models (x, b) \Leftrightarrow b = c(x)$.

Example 3 (Labeled Examples and Probabilistic Concepts) *Let $Z = X \times \{0, 1\}$ be again a set of labeled examples and let C be a family of functions from X to $[0, 1]$. Let the consistency relation be given by*

$$\forall c \in C, x \in X : c \models (x, 1) \Leftrightarrow c(x) > 0 \quad \textit{and} \quad c \models (x, 0) \Leftrightarrow c(x) < 1 \ .$$

Intuitively we should think of $c(x)$ as the probability that c assigns label 1 to instance x. If all concepts $c \in C$ were $0, 1$-valued, we would again be in the setting of Example 2.

Note that within Examples 1 and 2, we have that

$$\forall c, c' \in C : \ c \neq c' \Rightarrow Z_c \neq Z_{c'} \tag{3}$$

so that each concept $c \in C$ is uniquely determined by the full set Z_c of observations that c is consistent with. Of course this will not necessarily be the case if the concepts are probabilistic as in Example 3.

2.2 Variants of Sampling

As formalized in the definitions below, we distinguish between ordered and unordered sampling and we may have sampling with or without replacement.

Definition 1 (Sampling with Replacement) *Let $Q = (q(z))_{z \in Z}$ be a* collection of probability parameters, *i.e., $q(z) \geq 0$ and $\sum_{z \in Z} q(z) = 1$. For $n \geq 0$, we define n-*fold (ordered resp. unordered) Q-sampling with replacement *as the following random procedure:*

1. *Choose $z_1, \ldots, z_n$ independently at random according to Q.*
2. *In case of ordered sampling, return the sequence $(z_1, \ldots, z_n)$ whereas, in case of unordered sampling, return the multiset $\{z_1, \ldots, z_n\}$.*[3]

Let $\mathbf{z} = (z_1, \ldots, z_n) \in Z^n$ be a sequence that contains k distinct elements, say $z'_1, \ldots, z'_k$, and let n_i denote the number of occurrences of z'_i in $\mathbf{z}$. Let $A_{\mathbf{z}} \subseteq Z$ be the corresponding multiset. The probability that $\mathbf{z}$ (resp. $A_{\mathbf{z}}$) is obtained from n-fold ordered (resp. unordered) Q-sampling with replacement is henceforth denoted by $P^{O,R}(\mathbf{z}|Q)$ (resp. by $P^{\overline{O},R}(A_{\mathbf{z}}|Q)$). With these notations, the following holds:

$$P^{O,R}(\mathbf{z}|Q) = \prod_{i=1}^{n} q(z_i) = \prod_{i=1}^{k} q(z'_i)^{n_i} \quad \text{and} \quad P^{\overline{O},R}(A_{\mathbf{z}}|Q) = \frac{n!}{n_1! \ldots n_k!} \cdot \prod_{i=1}^{k} q(z'_i)^{n_i} \quad . \tag{4}$$

Definition 2 (Sampling without Replacement) *Let $Q = (q(z))_{z \in Z}$ be a collection of probability parameters. Let $N^+(Q)$ be the number of $z \in Z$ such that $q(z) > 0$. For $0 \leq n \leq N^+(Q)$, we define n-*fold (ordered resp. unordered) Q-sampling without replacement *as the following random procedure:*

[3] If $n = 0$, then the empty sequence resp. the empty multiset is returned,

1. *Choose z_1 at random according to Q.*
2. *For $i = 2, \ldots, n$ do the following:*
   Choose $z_i \in Z \setminus \{z_1, \ldots, z_{i-1}\}$ at random where, for each $z \in Z \setminus \{z_1, \ldots, z_{i-1}\}$, the probability for $z_i = z$ equals $\frac{q(z)}{1-(q(z_1)+\ldots+q(z_{i-1}))}$.[4]
3. *In case of ordered sampling, return the sequence $(z_1, \ldots, z_n)$ whereas, in case of unordered sampling, return the set $\{z_1, \ldots, z_n\}$.*

Let $\mathbf{z} = (z_1, \ldots, z_n) \in Z^n$ be a repetition-free sequence and let $A_{\mathbf{z}} \subseteq Z$ be the corresponding set. For a permutation σ of $1, \ldots, n$, we define $\mathbf{z}_\sigma = (z_{\sigma(1)}, \ldots, z_{\sigma(n)})$. The probability that $\mathbf{z}$ (resp. $A_{\mathbf{z}}$) is obtained from n-fold ordered (resp. unordered) Q-sampling without replacement is henceforth denoted by $P^{O,\overline{R}}(\mathbf{z}|Q)$ (resp. by $P^{\overline{O},\overline{R}}(A_{\mathbf{z}}|Q)$). With these notations, the following holds:

$$P^{O,\overline{R}}(\mathbf{z}|Q) = \prod_{i=1}^{n} \frac{q(z_i)}{1-(q(z_1)+\ldots+q(z_{i-1}))} \quad \text{and} \quad P^{\overline{O},\overline{R}}(A_{\mathbf{z}}|Q) = \sum_{\sigma} P^{O,\overline{R}}(\mathbf{z}_\sigma|Q) \ , \qquad (5)$$

where σ ranges over all permutations of $1, \ldots, n$.
We introduce the following notation:

- $\mathcal{Z}^{O,R} = Z^*$ denotes the set of sequences over Z (including the empty sequence).
- $\mathcal{Z}^{\overline{O},R}$ denotes the set of multisets over Z (including the empty multiset).
- $\mathcal{Z}^{O,\overline{R}}$ denotes the set of repetition-free sequences over Z (including the empty sequence).
- $\mathcal{Z}^{\overline{O},\overline{R}} = 2^Z$ denotes the powerset of Z.

The pairs $(\alpha, \beta) \in \{O, \overline{O}\} \times \{R, \overline{R}\}$ are called *sampling modes*. We use the symbol $\emptyset$ not only to denote the empty set but also to denote the empty multiset or the empty sequence. If A is a finite set or multiset, then $|A|$ denotes its size where, in case of a multiset, the multiple occurrences of elements are taken into account. The length of a finite sequence $\mathbf{z}$ is denoted by $|\mathbf{z}|$.

Remark 3 (Trivial Identities) *Suppose that $Q = (q(z))_{z \in Z}$ is collection of probability parameters. Then, for each sampling mode (α, β), we have that $P^{\alpha,\beta}(\emptyset|Q) = 1$. Moreover, if all parameters $q(z)$ with $z \in Z$ are strictly positive, then $P^{\overline{O},\overline{R}}(Z|Q) = 1$.*

We close this section with a more or less obvious result whose proof will be given for sake of completeness.

Remark 4 *Let $z_1, \ldots, z_n$ be a sequence with pairwise distinct elements from Z. Let $p_1 > p_2 > \ldots > p_n$ be a strictly decreasing sequence of strictly positive parameters such that $\sum_{i=1}^{n} p_i \leq 1$. For each permutation σ of $[n]$, consider the parameter collection $Q_\sigma = (q_\sigma(z_i))_{i=1,\ldots,n}$ given by $q_\sigma(z_i) = p_{\sigma(i)}$. Then the identity permutation is the unique maximizer of $P^{O,\overline{R}}(z_1, \ldots, z_n|Q_\sigma)$.*

[4] Note that the probability parameters for $z \in Z \setminus \{z_1, \ldots, z_{i-1}\}$ are the same as before up to normalization.

Proof According to (5), we have

$$
\begin{aligned}
P^{O,\overline{R}}(z_1,\ldots,z_k|Q_\sigma) &= \prod_{i=1}^{n} \frac{q_\sigma(z_i)}{1-(q_\sigma(z_1)+\ldots+q_\sigma(z_{i-1}))} \\
&= \prod_{i=1}^{n} \frac{p_{\sigma(i)}}{1-(p_{\sigma(1)}+\ldots+p_{\sigma(i-1)})} = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n}(1-(p_{\sigma(1)}+\ldots+p_{\sigma(i-1)})}
\end{aligned}
$$

The product in the numerator is the same for all permutations σ. The following assertions are equivalent:

1. σ^* is the identity permutation.
2. The sequence $p_{\sigma^*(1)},\ldots,p_{\sigma^*(n)}$ is strictly decreasing.
3. For each permutation $\sigma \neq \sigma^*$ and each $i \in [n]$, we have that

   $$p_{\sigma^*(1)}+\ldots+p_{\sigma^*(i-1)} \geq p_{\sigma(1)}+\ldots+p_{\sigma(i-1)}$$

   and, for at least one $i \in [n]$, this inequality is strict.
4. The permutation σ^* is the unique maximizer of $P^{O,\overline{R}}(z_1,\ldots,z_k|Q_\sigma)$.

The remark now is immediate from the equivalence of the first and the fourth statement. ■

2.3 MAP- and MLE-based Teaching

An MLE-learner will always choose a hypothesis from a class C that maximizes the likelihood of a given set of observations. MAP-learners are a bit more general because they additionally bring into play priors $(P(c))_{c\in C}$. The notion of likelihood depends on how the observations are randomly sampled. We proceed with the formal definition of MAP- and MLE-learners and their teachers:

Definition 5 (MAP- and MLE-Learner) *A* MAP-Learner L for C *is given by (and henceforth identified with) parameters* $P(z|c) \geq 0$ *and* $P(c) > 0$ *for* $z \in Z$ *and* $c \in C$ *such that*

$$\sum_{c\in C} P(c) = 1 \quad \text{and} \quad \forall c \in C : \sum_{z\in Z} P(z|c) = 1 \enspace .$$

The parameters $P(c)$ *are referred to as* priors*. The parameters* $P(z|c)$*, referred to as* c-conditional likelihoods*, must satisfy the following* validity condition*:*

$$c \not\models z \Rightarrow P(z|c) = 0 \enspace . \tag{6}$$

Set $Z_c^+(L) := \{z \in Z : P(z|c) > 0\}$ *and* $N^+(C,L) = \min_{c\in C}|Z_c^+(L)|$.[5] *$L$ can be in four different sampling modes (depending on the assumed kind of sampling). These modes determine the form of L's input and the choice of its output as will be detailed below.*

[5] Because of the validity condition, $Z_c^+(L)$ is a subset of $Z_c = \{z \in Z : c \models z\}$.

(O,R)**-mode:** *For every $n \geq 0$ and every sequence $\mathbf{a} \in Z^n$, we denote by $P^{O,R}(\mathbf{a}|c)$ the probability that $\mathbf{a}$ is obtained from n-fold ordered $P(\cdot|c)$-sampling with replacement. Given a sequence $\mathbf{a} \in \mathcal{Z}^{O,R}$, L returns the concept* $\mathrm{arg!max}_{c \in C}\left[P(c) \cdot P^{O,R}(\mathbf{a}|c)\right]$ *if it exists, and a question mark otherwise.*[6]

$(\overline{O},R)$**-mode:** *For every $n \geq 0$ and and every multiset $A \subseteq Z$ of size n, we denote by $P^{\overline{O},R}(A|c)$ the probability that A is obtained from n-fold unordered $P(\cdot|c)$-sampling with replacement. Given a multiset $A \in \mathcal{Z}^{\overline{O},R}$, L returns the concept* $\mathrm{arg!max}_{c \in C}\left[P(c) \cdot P^{\overline{O},R}(A|c)\right]$ *if it exists, and a question mark otherwise.*

$(O,\overline{R})$**-mode:** *For every $0 \leq n \leq N^+(C,L)$, and every repetition-free sequence $\mathbf{a} \in Z^n$, we denote by $P^{O,\overline{R}}(\mathbf{a}|c))$ the probability that $\mathbf{a}$ is obtained from n-fold ordered $P(\cdot|c)$-sampling without replacement. Given a repetition-free sequence $\mathbf{a} \in \mathcal{Z}^{O,\overline{R}}$ with $|\mathbf{a}| \leq N^+(C,L)$, L returns the concept* $\mathrm{arg!max}_{c \in C}\left[P(c) \cdot P^{O,\overline{R}}(\mathbf{a}|c)\right]$ *if it exists, and a question mark otherwise. If $|\mathbf{a}| > N^+(C,L)$, then also a question mark is returned.*

$(\overline{O},\overline{R})$**-mode:** *For every $0 \leq n \leq N^+(C,L)$, and every set $A \subseteq Z$ of size n, we denote by $P^{\overline{O},\overline{R}}(A|c)$ the probability that A is obtained from n-fold unordered $P(\cdot|c)$-sampling without replacement. Given a set $A \in \mathcal{Z}^{\overline{O},\overline{R}}$ with $|A| \leq N^+(C,L)$, L returns the concept* $\mathrm{arg!max}_{c \in C}\left[P(c) \cdot P^{\overline{O},\overline{R}}(A|c)\right]$ *if it exists, and a question mark otherwise. If $|A| > N^+(C,L)$, then also a question mark is returned.*

An MLE-learner *is a MAP-learner with uniform priors (so that the factor $P(c)$ in the above arg!max-expressions can be dropped).*

Definition 6 (Teacher) *Suppose that L is a MAP-learner for C that is in sampling mode $(\alpha,\beta) \in \{O,\overline{O}\} \times \{R,\overline{R}\}$. A* (successful) teacher *for L is a mapping T which assigns to each concept $c_0 \in C$ an input $I = T(c_0)$ for L such that $L(I) = c_0$. In other words:*

1. *$I \in \mathcal{Z}^{\alpha,\beta}$ and, if $\beta = \overline{R}$, then $|I| \leq N^+(C,L)$.*
2. *$c_0 = \mathrm{arg!max}_{c \in C}\left[P(c) \cdot P^{\alpha,\beta}(I|c)\right]$.*

A couple of observations are in place here.

Remark 7 *Suppose that L is a MAP-learner for C which is in sampling mode $(\alpha,\beta) \in \{O,\overline{O}\} \times \{R,\overline{R}\}$. Suppose that T is a teacher for L. Then the following holds for all $c,c' \in C$:*

$$L(T(c)) = c\ ,\ P^{\alpha,\beta}(\emptyset|c) = 1\ ,\ P^{\alpha,\beta}(T(c)|c) > 0\ ,\ c \models T(c)\ \ \textit{and}\ \ (c \neq c' \Rightarrow T(c) \neq T(c'))\ . \tag{7}$$

Moreover, if L is an MLE-learner and T is a teacher for L, then $T(c) \neq \emptyset$.

[6] The operator $\mathrm{arg!max}_{c \in C} f(c)$ returns the **unique** maximizer $c^* \in C$ of $f(c)$ provided that it exists.

Proof $L(T(c)) = c$ is an immediate consequence of Definitions 5 and 6. It now follows that, if $T(c) = T(c')$, then $c = L(T(c)) = L(T(c')) = c'$. In other words, $c \neq c'$ implies that $T(c) \neq T(c')$. 0-fold sampling conditioned to c yields $\emptyset$ regardless of how c is chosen. It follows that $P^{\alpha,\beta}(\emptyset|c) = 1$. Assume now for contradiction that $P^{\alpha,\beta}(T(c')|c') = 0$. But then c' cannot be the unique maximizer of $P^{\alpha,\beta}(T(c')|c) \cdot P(c)$ in C. This is in contradiction with $L(T(c')) = c'$. Assume for contradiction that $T(c)$ contains an observation $z \in Z$ such that $c \not\models z$. It follows that $P^{\alpha,\beta}(T(c)|c) = 0$, which is in contradiction with $P^{\alpha,\beta}(T(c)|c) > 0$. Thus $c \models T(c)$. Finally, suppose that the priors are uniform, i.e., $P(c) = 1/|C|$ for every $c \in C$. Assume for contradiction that $T(c_0) = \emptyset$ for some $c_0 \in C$. For every $c \in C$, we have $P(c) \cdot P^{\alpha,\beta}(\emptyset|c) = P(c) = 1/|C|$. Hence c_0 cannot be unique maximizer of $P(c) \cdot P^{\alpha,\beta}(\emptyset|c)$ in C. This is in contradiction with $L(T(c_0)) = c_0$. ■

Here is the definition of the parameter that is in the focus of our interest:

Definition 8 (MAP- and MLE-Teaching Dimension) *Suppose that L is a MAP-learner for C who is in sampling mode (α, β). The* MAP-teaching dimension of C given L and (α, β)*, denoted as* $\text{MAP-TD}_L^{\alpha,\beta}(C)$*, is defined as the smallest number d such that there exists a teacher of order d for L, respectively as ∞ if there does not exist a teacher for L. The* MAP-teaching dimension of C with respect to sampling mode (α, β) *is then given by*

$$\text{MAP-TD}^{\alpha,\beta}(C) := \min_L \text{MAP-TD}_L^{\alpha,\beta}(C) \ ,$$

where L ranges over all MAP-learners for C. Similarly, the MLE-teaching dimension of C with respect to sampling mode (α, β) *is given by* $\text{MLE-TD}^{\alpha,\beta}(C) := \min_L \text{MAP-TD}_L^{\alpha,\beta}(C)$ *with L ranging over all MLE-learners for C.*

The parameter $\text{MAP-TD}^{\alpha,\beta}(C)$ equals the number of observations needed to teach an optimally parameterized learner. It represents an information-theoretic barrier that cannot be broken regardless of how the learner is parameterized. Of course, this parameter will generally be smaller than the parameter $\text{MAP-TD}_L^{\alpha,\beta}(C)$ associated with a "naturally parameterized" learner. We close this section by mentioning the inequality

$$\text{MAP-TD}^{\alpha,\beta}(C) \leq \text{MLE-TD}^{\alpha,\beta}(C) \ ,$$

which (for trivial reasons) holds for each choice of C and (α, β).

3. Basic Results on the MAP-Based Teaching Model

In Ferri et al. (2022), the authors used a more restrictive condition in place of the validity condition. However, as we will see in Section 3.1, in the context of MAP-learners and their teachers, both conditions lead essentially to the same results. In Section 3.2, we discuss two natural monotonicity properties and thereafter, in Section 3.3, we note the equivalence of (O, R)- and the $(\overline{O}, R)$-mode and prove the pairwise incomparability of the modes $(\overline{O}, R)$, $(O, \overline{R})$ and $(\overline{O}, \overline{R})$.

3.1 Validity and Strong Validity

We will refer to

$$c \not\models z \Leftrightarrow P(z|c) = 0$$

as the *strong validity condition* for the parameters $(P(z|c))_{z\in Z,c\in C}$. This is the condition that the authors of Ferri et al. (2022) had imposed on the c-conditional likelihoods associated with a MAP-learner. We will see that each L satisfying the validity condition has a "close relative" L_ε that satisfies the strong validity condition. Here comes the definition of L_ε:

Definition 9 (ε-Shift) *Let L be given by parameters $P(c)$ and $P(z|c)$ with $c \in C$ and $z \in Z$ such that the validity condition is satisfied but the strong validity condition is not. We say that L_ε (with $0 < \varepsilon \le 1/2$) is the ε*-shift *of L if L_ε is given by the parameters $P(c)$ and $P_\varepsilon(z|c)$ where*

$$P_\varepsilon(z|c) = \begin{cases} (1-\varepsilon)\cdot P(z|c) & \text{if } z \in Z_c^+(L) \\ \frac{\varepsilon}{|Z_c \setminus Z_c^+(L)|} & \text{if } z \in Z_c \setminus Z_c^+(L) \\ 0 & \text{if } z \in Z \setminus Z_c \end{cases} .$$

For convenience, we set $P_\varepsilon(z|c) = P(z|c)$ if already L satisfies the strong validity condition.

Note that L_ε satisfies the strong validity condition because $P_\varepsilon(z|c) = 0$ iff $z \notin Z_c$ and $Z_c = \{z \in Z : c \models z\}$. A learner and its ε-shift are related as follows:

Lemma 10 *Let L be a MAP-learner for C whose parameters satisfy the validity condition. Then the following holds for each $(\alpha, \beta) \in \{O, \overline{O}\} \times \{R, \overline{R}\}$ and all sufficiently small $\varepsilon > 0$: each teacher for L in sampling mode (α, β) is also a teacher for L_ε in sampling mode (α, β).*

Proof Suppose that L and L_ε are both in sampling mode (α, β). Consider a teacher T for L. We claim that the following holds:

$$\forall c_0, c \in C : \lim_{\varepsilon\to 0} P_\varepsilon^{\alpha,\beta}(T(c_0)|c) = P^{\alpha,\beta}(T(c_0)|c) \ . \tag{8}$$

This would imply that, for every $c_0 \in C$ and sufficiently small ε, we have

$$c_0 = \text{arg!max}_{c\in C}\, P^{\alpha,\beta}(T(c_0)|c) = \text{arg!max}_{c\in C}\, P_\varepsilon^{\alpha,\beta}(T(c_0)|c) \ ,$$

which, in turn, implies that T is a teacher for L_ε. We still have to verify (8). This can be done by means of a simple continuity argument. Note first that

$$\forall c \in C, z \in Z : \lim_{\varepsilon\to 0} P_\varepsilon(z|c) = P(z|c) \ .$$

Since $P_\varepsilon^{\alpha,R}(T(c_0)|c)$ is a polynomial (and hence a continuous function) in the variables $P_\varepsilon(z|c)$ with $z \in T(c_0)$, we may conclude that (8) is true in case of $\beta = R$. Suppose now that $(\alpha, \beta) = (O, \overline{R})$ and $T(c_0) = (z_1, \ldots, z_n)$, which implies that $n \le N^+(C, L)$ and $z_1, \ldots, z_n \in Z_c^+(L)$. The function

$$P_\varepsilon^{O,\overline{R}}(T(c_0)|c) = \prod_{i=1}^{n} \frac{P_\varepsilon(z_i|c)}{1 - (P_\varepsilon(z_1|c) + \ldots + P_\varepsilon(z_{i-1}|c))}$$

is a rational function in the variables $P_\varepsilon(z_i|c)$ for $i = 1, \ldots, n$. Hence we can apply the continuity argument again but, in addition, we must rule out that the denominator, $1 - (P_\varepsilon(z_1|c) + \ldots + P_\varepsilon(z_{i-1}|c))$, converges to 0 when ε approaches 0. This, however, can be ruled out as follows:

- Set $\rho := \frac{1}{2} \cdot \min_{c \in C, z \in Z_c^+(L)} P(z|c)$ and note that $0 < \rho \leq \min_{c \in C, z \in Z_c^+(L)} P_\varepsilon(z|c)$. The latter inequality holds because of $P_\varepsilon(z|c) = (1 - \varepsilon) \cdot P(z|c)$ and $\varepsilon \leq 1/2$.
- Because of $n \leq N^+(C, L)$, the set $\{z_1, \ldots, z_{n-1}\}$ cannot contain all elements of $Z_c^+(L)$.
- Therefore $1 - (P_\varepsilon(z_1|c) + \ldots + P_\varepsilon(z_{i-1}|c) \geq \rho$ for all $i = 1, \ldots, n$ and the limit for $\varepsilon \rightarrow 0$ cannot be equal to 0.

We may therefore conclude that (8) is true in case of $(\alpha, \beta) = (O, \overline{R})$. The proof in case of $(\alpha, \beta) = (\overline{O}, \overline{R})$ is similar. ■

Corollary 11 *With the notation from Definition 9, we have*

$$\text{MAP-TD}_L^{\alpha,\beta}(C) = \text{MAP-TD}_{L_\varepsilon}^{\alpha,\beta}(C)$$

for all sufficiently small $\varepsilon > 0$.

3.2 Monotonicity Properties

It is clear, intuitively, that adding concepts without adding observations should make the teaching problem harder. Conversely, adding observations without adding concepts should make the teaching problem easier. In this section, we formalize these statements and prove them. All results in this section are formulated in terms of MAP-TD. But the corresponding results with MLE-TD in place of MAP-TD hold as well.

We say that $(C', Z', \models')$ is an *extension* of $(C, Z, \models)$ if $C \subseteq C'$, $Z \subseteq Z'$ and, for all $c \in C$ and $z \in Z$, we have that $c \models z$ if and only if $c \models' z$.

So far, we used a notation (e.g. $\text{MAP-TD}^{\alpha,\beta}(C)$ instead of $\text{MAP-TD}^{\alpha,\beta}(C, Z, \models)$) which made a dependence on $(C, Z, \models)$ explicit for C only (because the corresponding Z and the corresponding relation $\models$ were clear from context). In this section, there is some danger of confusion and, consequently, we use a notation which makes the dependence on the whole triple $(C, Z, \models)$ more explicit.

Definition 12 *Let $(C', Z', \models')$ be an extension of $(C, Z, \models)$ with $Z' = Z$. Let L be a MAP-learner for $(C', Z, \models')$ with parameters $P(c') > 0$ and $P(z|c')$ for $c' \in C'$ and $z \in Z$. Set $P(C) = \sum_{c \in C} P(c)$. The MAP-learner with parameters $P(c)/P(C)$ and $P(z|c)$ for $c \in C$ and $z \in Z$, denoted by $L_{\downarrow C}$, is called the* restriction of L to subclass C.

The parameters of a MAP-learner L for $(C', Z, \models')$ must satisfy the validity condition. Clearly the parameters of $L_{\downarrow C}$ satisfy the validity condition too. Moreover, for each $c \in C$, we have that $Z_c^+(L_{\downarrow C}) = Z_c^+(L)$. These observations can be used for showing the following result:

Lemma 13 (Concept-Class Monotonicity) *With the assumptions and notation as in Definition 12, the following holds for each sampling mode (α, β):*

$$\text{MAP-TD}^{\alpha,\beta}_{L_{\downarrow C}}(C, Z, \models) \ \le\ \text{MAP-TD}^{\alpha,\beta}_{L}(C', Z, \models') \ .$$

Proof Let $T : C' \rightarrow \mathcal{Z}^{\alpha,\beta}$ be a teacher for L and let $T_{\downarrow C}$ denote its restriction to subclass C. Clearly the order of $T_{\downarrow C}$ is upper-bounded by the order of T. It suffices to show that $T_{\downarrow C}$ is a teacher for $L_{\downarrow C}$. To this end, we have to show the following:

(a) If $\beta = \overline{R}$ then, for all $c \in C$, we have that $|T_{\downarrow C}(c)| \le N^+(C, L_{\downarrow C})$.

(b) For all $c_0 \in C$, $c \in C \backslash \{c_0\}$, we have that $P(c) \cdot P^{\alpha,\beta}(T_{\downarrow C}(c_0)|c) < P(c_0) \cdot P^{\alpha,\beta}(T_{\downarrow C}(c_0)|c_0)$.

Of course, since T is teacher for L, we know that the following hold:

(a') If $\beta = \overline{R}$ then, for all $c' \in C'$, we have that $|T(c')| \le N^+(C', L)$.

(b') For all $c'_0 \in C'$, $c' \in C' \backslash \{c'_0\}$, we have that $P(c') \cdot P^{\alpha,\beta}(T(c'_0)|c') < P(c'_0) \cdot P^{\alpha,\beta}(T(c'_0)|c'_0)$.

The following calculation verifies (a) under the assumption that $\beta = \overline{R}$:

$$\begin{aligned} |T_{\downarrow C}(c)| \ &=\ |T(c)| \ \le\ N^+(C', L) \ =\ \min_{c' \in C'} |Z^+_{c'}(L)| \\ &\le\ \min_{c \in C} |Z^+_c(L)| \ =\ \min_{c \in C} |Z^+_c(L_{\downarrow C})| \ =\ N^+(C, L_{\downarrow C}) \ . \end{aligned}$$

Suppose that $c_0 \in C$ and $c \in C \setminus \{c_0\}$. Then (b) can be verified as follows:

$$P(c) \cdot P^{\alpha,\beta}(T_{\downarrow C}(c_0)|c) = P(c) \cdot P^{\alpha,\beta}(T(c_0)|c) < P(c_0) \cdot P^{\alpha,\beta}(T(c_0)|c_0) = P(c_0) \cdot P^{\alpha,\beta}(T_{\downarrow C}(c_0)|c_0) \ .$$

Here the first and the last equation hold because $c_0 \in C$ and therefore $T_{\downarrow C}(c_0) = T(c_0)$. ■

Corollary 14 *If $(C', Z', \models')$ is an extension of $(C, Z, \models)$ with $Z = Z'$, then*

$$\text{MAP-TD}^{\alpha,\beta}(C, Z, \models) \ \le\ \text{MAP-TD}^{\alpha,\beta}(C', Z, \models') \ .$$

Definition 15 *Let $(C', Z', \models')$ be an extension of $(C, Z, \models)$ with $C' = C$. Let L be a MAP-learner for $(C, Z, \models)$ with parameters $P(c)$ and $P(z|c)$ for $c \in C$ and $z \in Z$. The MAP-learner with parameters $P_{\uparrow Z'}(c) = P(c)$ and*

$$P_{\uparrow Z'}(z'|c) = \begin{cases} P(z'|c) & \text{if } z' \in Z \\ 0 & \text{otherwise} \end{cases} \ ,$$

denoted by $L_{\uparrow Z'}$, is called the extension of L to superset Z'.

The parameters of a MAP-learner L for $(C, Z, \models)$ must satisfy the validity condition. It is easy to check that, therefore, the parameters of $L_{\uparrow Z'}$ satisfy the validity condition too. Moreover, for each $c \in C$, we have that

$$\{z' \in Z' : P_{\uparrow Z'}(z'|c) > 0\} \ =\ \{z \in Z : P(z|c) > 0\} \ =\ Z^+_c(L) \ ,$$

which implies that $N^+(C, L_{\uparrow Z'}) = N^+(C, L)$. These observations can be used for showing the following result:

Lemma 16 (Observation-Set Monotonicity) *With the assumptions and the notation as in Definition 15, the following holds for each sampling mode (α, β):*

$$\text{MAP-TD}_L^{\alpha,\beta}(C, Z, \models) \;\geq\; \text{MAP-TD}_{L\uparrow Z'}^{\alpha,\beta}(C, Z', \models') \enspace .$$

Proof Let $T : C \to \mathcal{Z}^{\alpha,\beta}$ be a teacher for L. It is sufficient to show that T is also a teacher for $L_{\uparrow Z'}$ (albeit a teacher for $L_{\uparrow Z'}$ who does not make use of observations in $Z' \setminus Z$). To this end, we have to show the following:

(a) If $\beta = \overline{R}$ then, for all $c \in C$, we have that $|T(c)| \leq N^+(C, L_{\uparrow Z'})$.

(b) For all $c_0 \in C$, $c \in C \setminus \{c_0\}$, we have that $P(c) \cdot P_{\uparrow Z'}^{\alpha,\beta}(T(c_0)|c) < P(c_0) \cdot P_{\uparrow Z'}^{\alpha,\beta}(T(c_0)|c_0)$.

Assertion (a), assuming $\beta = \overline{R}$, is obtained by

$$|T(c)| \leq N^+(C, L) \;=\; N^+(C, L_{\uparrow Z'}) \enspace ,$$

where the first inequality holds because T is a teacher for L. Suppose that $c_0 \in C$ and $c \in C \setminus \{c_0\}$. Assertion (b) is obtained by

$$P(c) \cdot P_{\uparrow Z'}^{\alpha,\beta}(T(c_0)|c) = P(c) \cdot P^{\alpha,\beta}(T(c_0|c) < P(c_0) \cdot P^{\alpha,\beta}(T(c_0)|c_0) = P(c_0) \cdot P_{\uparrow Z'}^{\alpha,\beta}(T(c_0)|c_0) \enspace ,$$

where the first and the last equation holds because $T(c_0) \subseteq Z$ so that the likelihoods of observations in $Z' \setminus Z$ do not come into play. The inequality in the middle holds because T is a teacher for L. ■

Corollary 17 *If $(C', Z', \models')$ is an extension of $(C, Z, \models)$ with $C = C'$, then*

$$\text{MAP-TD}^{\alpha,\beta}(C, Z, \models) \;\geq\; \text{MAP-TD}^{\alpha,\beta}(C, Z', \models') \enspace .$$

3.3 A Comparison of the Sampling Modes

We say that the sampling mode (α, β) *dominates* the sampling mode (α', β') if, for every concept class C and every MAP-learner L for C, we have that $\text{MAP-TD}_L^{\alpha,\beta}(C) \leq \text{MAP-TD}_L^{\alpha',\beta'}(C)$. We say they are *equivalent* if they mutually dominate each other, i.e., if $\text{MAP-TD}_L^{\alpha,\beta}(C) = \text{MAP-TD}_L^{\alpha',\beta'}(C)$ holds for every choice of C and L. We say, they are *incomparable* if none of them dominates the other one. We start with an easy observation:

Remark 18 *The sampling modes (O, R) and $(\overline{O}, R)$ are equivalent.*

Proof Consider a concept class C and a MAP-learner L for C. Let $\mathbf{a} \in Z^n$ be a sequence of k distinct elements with multiplicities $n_1, \ldots, n_k$, respectively. Denote by A the corresponding multiset. An inspection of (4) shows that the following holds for each $c \in C$:

$$P^{\overline{O},R}(A|c) = \frac{n!}{n_1! \ldots n_k!} \cdot P^{O,R}(\mathbf{a}|c) \enspace . \tag{9}$$

Let $\mathbf{a}'$ be a sequence obtained from $\mathbf{a}$ by a permutation of the components. Since $\mathbf{a}'$ also consists of k distinct elements with multiplicities $n_1, \ldots, n_k$, respectively, equation (9) also holds with $\mathbf{a}'$ in place of $\mathbf{a}$. It therefore easily follows that a teacher T for L, with L being in sampling mode (O, R), can be converted into a teacher T' of the same order for L with L being in sampling mode $(\overline{O}, R)$, and vice versa:

- Suppose that T is given. If $T(c) = \mathbf{a}$, then define $T'(c) = A$ where A is the multiset induced by $\mathbf{a}$.
- Suppose that T' is given. If $T'(c) = A$ then define $T(A) = \mathbf{a}$ where $\mathbf{a}$ is an (arbitrarily chosen) sequence containing the same elements as A with the same multiplicities.

It follows from this discussion that $\text{MAP-TD}_L^{O,R}(C) = \text{MAP-TD}_L^{\overline{O},R}(C)$, which concludes the proof. ■

Corollary 19 $\text{MAP-TD}^{O,R}(C) = \text{MAP-TD}^{\overline{O},R}(C)$ *and* $\text{MLE-TD}^{O,R}(C) = \text{MLE-TD}^{\overline{O},R}(C)$.

We now turn our attention to the incomparability results:

Theorem 20 *The sampling modes* (O, R), $(O, \overline{R})$ *and* $(\overline{O}, \overline{R})$ *are pairwise incomparable.*

In order to prove the theorem, we will consider triples $(C, Z, \models)$ with $C = \{c_1, c_2, c_3\}$, $Z = \{z_1, z_2, z_3\}$ and $c_i \models z_j$ for all $1 \le i, j \le 3$. An important role will be played by concepts of the form $c^{\pm\Delta}$ with parameters given by

$$P(z_1|c^{\pm\Delta}) = p + \Delta \ , \ P(z_2|c^{\pm\Delta}) = p - \Delta \ \text{ and } \ P(z_3|c^{\pm\Delta}) = 1 - 2p \ . \tag{10}$$

The following Facts 1–4, which pave the way for the proof of Theorem 20, can be proven by using the derivation rules of analysis. For sake of completeness, these proofs are given in the appendix.

Fact 1: Suppose that $0 \le |\Delta| < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then $P^{O,R}(z_1, z_2|c^{\pm\Delta})$ and $P^{\overline{O},R}(z_1, z_2|c^{\pm\Delta})$ are both strictly decreasing when $|\Delta|$ is increased, which implies that $\Delta = 0$ is the unique maximizer.

Fact 2: Suppose that $0 \le |\Delta| < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then

$$P^{O,\overline{R}}(z_1, z_2|c^{\pm\Delta}) - P^{O,\overline{R}}(z_1, z_2|c^{\pm 0}) \begin{cases} = 0 & \text{if } \Delta \in \{0, \frac{p^2}{1-p}\} \\ > 0 & \text{if } 0 < \Delta < \frac{p^2}{1-p} \\ < 0 & \text{otherwise} \end{cases} . \tag{11}$$

Fact 3: Suppose that $0 \le \Delta < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then

$$P^{O,R}(z_1, z_1, z_2|c^{\pm\Delta}) - P^{O,R}(z_1, z_1, z_2|c^{\pm 0}) \begin{cases} = 0 & \text{if } \Delta \in \left\{0, \frac{1}{2}(\sqrt{5} - 1)p\right\} \\ > 0 & \text{if } 0 < \Delta < \frac{1}{2}(\sqrt{5} - 1)p \\ < 0 & \text{otherwise} \end{cases} . \tag{12}$$

Fact 4: Suppose that $0 < p < 1/2$ and $1 \le t < \frac{1-p}{p}$. Let $c^{(t)}$ be the concept whose parameters are given by

$$c^{(t)}(z_1) = pt \ , \ c^{(t)}(z_2) = p/t \ \text{ and } \ c^{(t)}(z_3) = 1 - pt - p/t \ . \tag{13}$$

Then $P^{\overline{O},\overline{R}}(z_1, z_2|c^{(t)})$ is strictly increasing with t.

A couple of more intuitive remarks are in place here. Fact 1 tells us that, in sampling modes (O, R) and $(\overline{O}, \overline{R})$, the better a concept explains observations z_1, z_2 (in the maximum likelihood sense), the more evenly it splits the available probability mass $2p$ among them. We will refer to an application of Fact 1 as applying the "even-split argument". In sampling mode $(O, \overline{R})$, however, the even split does not maximize the likelihood of these observations. The likelihood of z_1, z_2 becomes larger if the probability assigned to z_1 is slightly larger than the probability assigned to z_2. See (11). A similar remark applies to the sampling mode (O, R) and the sequence z_1, z_1, z_2. See (12). Fact 4 is concerned with sampling mode $(\overline{O}, \overline{R})$ and a multiplicative decomposition of p^2 into pt (the probability assigned to z_1) and p/t (the probability assigned to z_2) with $t \ge 1$. According to Fact 4, the likelihood of $\{z_1, z_2\}$ becomes larger when the scaling factor $t \ge 1$ is increased. Note that this is not in contradiction with the even-split argument, because $pt + p/t$ is itself strictly increasing with t so that the even-split argument does not apply.

We would furthermore like to note that the c-conditional likelihood of a (multi-)set or sequence of observations becomes larger if one of the relevant c-conditional likelihood parameters is increased while the others are fixed. We refer to this way of arguing as applying the "monotonicity argument".
Theorem 20 is a direct consequence of the following three lemmas.

Lemma 21 *Consider the triple* $(C, Z, \models)$ *with* $C = \{c_1, c_2, c_3\}$, $Z = \{z_1, z_2, z_3\}$ *and* $c_i \models z_j$ *for all* $1 \le i, j \le 3$. *Let* L *be an MLE-learner for* C *with parameters given by*

$P(z\|c)$	c_1	c_2	c_3
z_1	$p + \Delta_1$	$p + \Delta_2$	p
z_2	$p - \Delta_1$	$p - \Delta_2$	p
z_3	$1 - 2p$	$1 - 2p$	$1 - 2p$

,

where $0 < \Delta_1 < \frac{p^2}{1-p} < \Delta_2 = \frac{1}{2}(\sqrt{5} - 1)p < p \le 0.4$.[7] *Then*

$$\text{MLE-TD}_L^{O,R}(C) = 3 \ , \ \text{MLE-TD}_L^{O,\overline{R}}(C) = 2 \ \textit{ and } \ \text{MLE-TD}_L^{\overline{O},\overline{R}}(C) = \infty \ . \tag{14}$$

Proof It is obvious that, in any mode of sampling, the concept c_2 can be taught by observation z_1 and the concept c_3 can be taught by observation z_2. An inspection of (11) and (12) reveals that

$$\begin{aligned} P_L^{O,\overline{R}}(z_1, z_2|c_1) &> P_L^{O,\overline{R}}(z_1, z_2|c_3) > P_L^{O,\overline{R}}(z_1, z_2|c_2) \ , \\ P_L^{O,R}(z_1, z_1, z_2|c_1) &> \max\{P_L^{O,R}(z_1, z_1, z_2|c_2), P_L^{O,R}(z_1, z_1, z_2|c_3)\} \ . \end{aligned}$$

[7] The constraint $p \le 0.4$ has the effect that $\frac{p}{1-p} < \frac{1}{2}(\sqrt{5} - 1)$.

It follows that c_1 can be taught in $(O,\overline{R})$-mode (resp. in (O,R)-mode) by the sequence z_1, z_2 (resp. by the sequence z_1, z_1, z_2). We will argue now that there are no shorter sequences for teaching c_1 and that, in $(\overline{O},\overline{R})$-mode, c_1 cannot be taught at all. An application of the monotonicity argument yields that c_1 cannot be taught by a single observation (regardless of the sampling mode). The same remark holds for 2 observations except, possibly, for observations z_1, z_2. But, by the even-split argument, it is the concept c_3 that assigns the highest probability to the sequence $(z_1, z_2) \in \mathcal{Z}^{O,R}$ resp. to the set $\{z_1, z_2\} \in \mathcal{Z}^{\overline{O},\overline{R}}$. Thus $(O,\overline{R})$ is the only sampling mode in which c_1 can be taught by 2 observations. It follows that, in $(\overline{O},\overline{R})$-mode, c_1 cannot be taught at all.[8] We may conclude from this discussion that the identities in (14) are valid. ■

Lemma 21 implies that (O,R) does not dominate $(O,\overline{R})$ and $(\overline{O},\overline{R})$ does not dominate any of the other sampling modes. The next result leads to some more no-domination results:

Lemma 22 *Consider the triple $(C, Z, \models)$ with $C = \{c_1, c_2, c_3\}$, $Z = \{z_1, z_2, z_3\}$ and $c_i \models z_j$ for all $1 \le i, j \le 3$. Let L be an MLE-learner for C with the parameters $P(z|c)$ given by*

$P(z\|c)$	c_1	c_2	c_3
z_1	p	$p+\Delta$	$p-\Delta$
z_2	p	$p-\Delta$	$p+\Delta$
z_3	$1-2p$	$1-2p$	$1-2p$

,

where $0 < \Delta < \frac{p^2}{1-p} < p < 1/2$. Then

$$\text{MLE-TD}_L^{O,R}(C) = \text{MLE-TD}_L^{\overline{O},\overline{R}}(C) = 2 \quad \textit{and} \quad \text{MLE-TD}_L^{O,\overline{R}}(C) = \infty \ . \tag{15}$$

Proof Clearly the concept c_2 can be taught by observation z_1 and the concept c_3 can be taught by observation z_2 in any mode of sampling. The concept c_1 cannot be taught by a single observation. But it can be taught by the sequence (z_1, z_2) in (O,R)-mode and by the set $\{z_1, z_2\}$ in $(\overline{O},\overline{R})$-mode (application of the even-split argument). We finally discuss teachability of c_1 in $(O,\overline{R})$-mode. An application of the monotonicity argument yields that c_1 cannot be taught in $(O,\overline{R})$-mode by two observations except, possibly, by the observations (z_1, z_2) or (z_2, z_1) in $\mathcal{Z}^{O,\overline{R}}$. But an inspection of (11) reveals that it is the concept c_2 (resp. c_3) that assigns the highest probability to (z_1, z_2) (resp. to (z_2, z_1)). It follows that, in $(O,\overline{R})$-mode, the concept c_1 cannot be taught at all. We may conclude from this discussion that the identities in (15) are valid. ■

Lemma 22 implies that $(O,\overline{R})$ does not dominate any of the other sampling modes. The next result implies (O,R) does not dominate $(\overline{O},\overline{R})$.

[8] Here we make use of the fact that, if $Z_c = Z$ for each $c \in C$, then $P^{\overline{O},\overline{R}}(Z|c) = 1$ for each $c \in C$. Note that this rules out the possibility of having teaching sets of size $3 = |Z|$.

Lemma 23 *Consider the triple* $(C, Z, \models)$ *with* $C = \{c_1, c_2, c_3\}$, $Z = \{z_1, z_2, z_3\}$ *and* $c_i \models z_j$ *for all* $1 \leq i, j \leq 3$. *Let* L *be an MLE-learner for* C *with parameters* $P(z|c)$ *given by*

$P(z\|c)$	c_1	c_2	c_3
z_1	sp	p	$sp + \varepsilon$
z_2	p/s	p	$p/s - \varepsilon$
z_3	$1 - sp - p/s$	$1 - 2p$	$1 - sp - p/s$

,

where $0 < p < \frac{1}{2}$, $1 < s \leq \frac{1-p}{p}$ *and* $0 < \varepsilon < \min\{1 - sp, p/s\}$. *Then*

$$\text{MLE-TD}_L^{\overline{O},\overline{R}}(C) = 2 < \text{MLE-TD}_L^{O,R}(C) \ . \tag{16}$$

Proof Clearly, the concept c_2 can be taught by observation z_2 and c_3 can be taught by observation z_1 in any mode of sampling. It is obvious that c_1 cannot be taught by a single observation (regardless of the sampling mode). In (O, R)-mode, the concept c_1 cannot be taught by sequences of length 2 because c_1 is for none of them the unique maximizer:

- $P_L^{O,R}(z_1, z_2|c_1) = p^2 = P_L^{O,R}(z_1, z_2|c_2)$.
- $P_L^{O,R}(z_1, z_3|c_1) < P_L^{O,R}(z_1, z_3|c_3)$ and $P_L^{O,R}(z_2, z_3|c_1) < P_L^{O,R}(z_2, z_3|c_2)$.[9]

However, in $(\overline{O}, \overline{R})$-mode, the concept c_1 can be taught by the set $\{z_1, z_2\}$:

- Concept c_1 distributes the probability mass $sp + p/s$ (slightly) more evenly on z_1 and z_2 than the concept c_3. By the even-split argument, we obtain $P^{\overline{O},\overline{R}}(\{z_1, z_2\}|c_1) > P^{\overline{O},\overline{R}}(\{z_1, z_2\}|c_3)$.
- Recall from Fact 4 that $c^{(t)}$, with $t \geq 1$, denotes the concept which assigns probability pt to z_1, probability p/t to z_2 and the remaining probability mass to z_3. Note that $c_1 = c^{(s)}$ and $c_2 = c^{(1)}$. According to Fact 4, the function $P^{\overline{O},\overline{R}}(z_1, z_2|c^{(t)})$ is strictly increasing with t. Hence $P^{\overline{O},\overline{R}}(\{z_1, z_2\}|c_1) > P^{\overline{O},\overline{R}}(\{z_1, z_2\}|c_2)$.

The identities in (16) are immediate from this discussion. ■

Putting the above three lemmas together, we obtain Theorem 20.

4. MAP-Based Teaching and Saturating Matchings

Suppose that C is a concept class with observation set Z and consistency relation $\models$. The bipartite graph $G(C) = (C, Z, E)$ with

$$E = \{(c, z) \in C \times Z : c \models z\}$$

is called the *consistency graph (associated with C)*. Let $\mathcal{Z}^{\alpha,\beta}$ with $(\alpha, \beta) \in \{O, \overline{O}\} \times \{R, \overline{R}\}$ be the notation that was introduced in Section 2.2. The bipartite graph $G(C)^{\alpha,\beta} = (C, \mathcal{Z}^{\alpha,\beta}, E^{\alpha,\beta})$ with

$$E^{\alpha,\beta} = \{(c, \zeta) \in C \times \mathcal{Z}^{\alpha,\beta} : c \models \zeta\}$$

[9] These are two applications of the monotonicity argument. Note that $s + \frac{1}{s} > 2$ for all $s > 1$.

is called the *extended consistency graph (associated with C)*. The graph resulting from $G(C)^{\alpha,\beta}$ by the removal of the vertex $\emptyset$ from the second vertex class $\mathcal{Z}^{\alpha,\beta}$ will be denoted by $G(C)^{\alpha,\beta}_{\neq\emptyset}$. We denote by $\mathrm{SMN}(G(C)^{\alpha,\beta})$ the smallest possible order of a C-saturating matching in $G(C)^{\alpha,\beta}$. Analogously, $\mathrm{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset})$ denotes the smallest possible order of a C-saturating matching in $G(C)^{\alpha,\beta}_{\neq\emptyset}$. For ease of later reference, we make the following observation:

Remark 24 *Suppose that $T : C \rightarrow \mathcal{Z}^{\alpha,\beta}$ is a mapping which satisfies*

$$\forall c, c' \in C : (c \models T(c)) \wedge (c \neq c' \Rightarrow T(c) \neq T(c')) \ . \tag{17}$$

Then T is of order at least $\mathrm{SMN}(G(C)^{\alpha,\beta})$. Moreover, if T satisfies (17) and $\emptyset$ is not in the image of T, then T is of order at least $\mathrm{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset})$.

Proof If T satisfies (17), then T represents a C-saturating matching in $G(C)^{\alpha,\beta}$. If additionally $\emptyset$ is not in the image of T, then T represents a C-saturating matching in $G(C)^{\alpha,\beta}_{\neq\emptyset}$. ■

Here is the main result of this section:

Theorem 25 *For each sampling mode (α, β), we have*

$$\text{MAP-TD}^{\alpha,\beta}(C) \geq \mathrm{SMN}(G(C)^{\alpha,\beta}) \ \text{ and } \ \text{MLE-TD}^{\alpha,\beta}(C) \geq \mathrm{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset}) \ . \tag{18}$$

Moreover, for $(\alpha, \beta) = (\overline{O}, R)$, this holds with equality.

Proof Let L be a MAP-learner for C and let (α, β) denote its sampling mode. Let T be a teacher for L. Recall from (7) that T satisfies (17). Moreover, if L is an MLE-learner for C, then $T(c) \neq \emptyset$ for all $c \in C$. Now an application of Remark 24 yields (18).

We move on and prove that $\text{MLE-TD}^{\overline{O},R}(C) \leq \mathrm{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset})$. Suppose that M is a C-saturating matching in $G(C)^{\overline{O},R}_{\neq\emptyset}$ that is of order $\mathrm{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset})$. For each $c \in C$ and $z \in Z$, let $n(z, c)$ denote the number of occurrences of z in the multiset $M(c)$ and let $n(c) = |M(c)|$. Consider a learner L with uniform priors (= MLE-learner) and the parameters $P(z|c) = \frac{n(z,c)}{n(c)}$. Note that these parameters satisfy the validity condition. It suffices to show that M represents a teacher for L, i.e., we have to show that

$$\forall c^* \in C : c^* = \mathrm{arg!max}_{c \in C}\, P^{\overline{O},R}(M(c^*)|c) \ .$$

To this end, we pick a concept c from $C \setminus \{c^*\}$, and proceed by case analysis:

Case 1: $M(c^*)$ and $M(c)$ contain the same elements of Z (albeit with different multiplicities)[10].

Denote these elements by $z_1, \ldots, z_k$. Let $n := n(c^*)$, $n_i = n(z_i, c^*)$. Then $p_i := n_i/n$

[10] The multiplicities cannot be the same because $M : C \rightarrow \mathcal{Z}^{\overline{O},R}$ is a matching.

is the relative frequency of z_i in $M(c^*)$. Let q_i denote the relative frequency of z_i in $M(c)$, which implies that $\mathbf{q} \neq \mathbf{p}$. It follows that

$$P^{\overline{O},R}(M(c^*)|c^*) = \frac{n!}{n_1! \ldots n_k!} \cdot \prod_{i=1}^{k} p_i^{n_i} \text{ and } P^{\overline{O},R}(M(c^*)|c) = \frac{n!}{n_1! \ldots n_k!} \cdot \prod_{i=1}^{k} q_i^{n_i} \ .$$

A straightforward calculation shows that $P^{\overline{O},R}(M(c^*)|c^*) > P^{\overline{O},R}(M(c^*)|c)$ iff

$$\sum_{i=1}^{k} p_i \log\left(\frac{p_i}{q_i}\right) > 0 \ . \tag{19}$$

The left-hand side is the Kullback-Leibler divergence (= KLD) between $\mathbf{p}$ and $\mathbf{q}$. Since the KLD is non-negative and 0 only if $\mathbf{q} = \mathbf{p}$, the condition (19) is satisfied.

Case 2: $M(c^*)$ contains an element that is not contained in $M(c)$.
Then the c-conditional likelihood of $M(c^*)$ equals 0.

Case 3: All elements in $M(c^*)$ are contained in $M(c)$, but $M(c)$ contains an element that is not contained in $M(c^*)$.
Then the c-conditional likelihood of $M(c^*)$ can be expressed as $\Pr(E_1) \cdot \Pr(E_2|E_1)$ for the following two events:

E_1: $n(c^*)$-fold c-sampling yields only elements from $M(c^*)$.

E_2: $n(c^*)$-fold c-sampling yields $M(c^*)$.

Since $M(c)$ contains an element that is not contained in $M(c^*)$, we have $\Pr(E_1) < 1$. It follows from the analysis of Case 1 that $\Pr(E_2|E_1)$ is upper-bounded by the c^*-conditional likelihood of $M(c^*)$.

We may conclude from the above discussion that $c^* = \text{arg!max}_{c \in C} P^{\overline{O},R}(M(c^*)|c)$. Thus M can be seen as a teacher for L. It follows that $\text{MLE-TD}^{\overline{O},R}(C) \leq \text{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset})$.
The inequality $\text{MAP-TD}^{\overline{O},R}(C) \leq \text{SMN}(G(C)^{\overline{O},R})$ can be obtained in a similar fashion. We start with a C-saturating matching M in $G(C)^{\overline{O},R}$ that is of order $\text{SMN}(G(C)^{\overline{O},R})$. If M does not assign $\emptyset$ to any concept, we can proceed as before. Otherwise, if $M(c_0) = \emptyset$ for some $c_0 \in C$, we still use a similar reasoning but with a slight modification of the parameter collection of the learner L:

- The priors are given by setting $P(c_0) = \frac{1+\varepsilon}{|C|}$ and by letting the remaining $|C| - 1$ concepts evenly share the remaining probability mass (still almost uniform priors).
- The parameters $P(z|c)$ are chosen as before.

We can again view the matching M as a teacher for L. Since $P^{\overline{O},R}(\emptyset|c) = 1$ for all $c \in C$, we obtain

$$\text{arg!max}_{c \in C} \left(P(c) \cdot P^{\overline{O},R}(\emptyset|c)\right) = \text{arg!max}_{c \in C} P(c) = c_0 \ .$$

For the remaining concepts, the reasoning is as before provided that $\varepsilon > 0$ s sufficiently small: this is an easy continuity argument which exploits that the priors converge to the uniform distribution on C if ε approaches 0. ■

Clearly

$$\begin{aligned} \mathrm{SMN}(G(C)^{O,R}) &\le \min\{\mathrm{SMN}(G(C)^{\overline{O},R}),\mathrm{SMN}(G(C)^{O,\overline{R}})\} \\ &\le \max\{\mathrm{SMN}(G(C)^{\overline{O},R}),\mathrm{SMN}(G(C)^{O,\overline{R}})\} \le \mathrm{SMN}(G(C)^{\overline{O},\overline{R}}) \end{aligned}$$

and

$$\begin{aligned} \mathrm{SMN}(G(C)^{O,R}_{\neq\emptyset}) &\le \min\{\mathrm{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset}),\mathrm{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})\} \\ &\le \max\{\mathrm{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset}),\mathrm{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})\} \le \mathrm{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset}) \ . \end{aligned}$$

Combining this with Theorem 25 and with Corollary 19, we immediately obtain the following result:

Corollary 26

1. $\text{MAP-TD}^{\overline{O},R}(C) = \mathrm{SMN}(G(C)^{\overline{O},R}) \le \mathrm{SMN}(G(C)^{\overline{O},\overline{R}}) \le \text{MAP-TD}^{\overline{O},\overline{R}}(C)$.

2. $\text{MLE-TD}^{\overline{O},R}(C) = \mathrm{SMN}(G(C)^{\overline{O},R}_{\neq\emptyset}) \le \mathrm{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset}) \le \text{MLE-TD}^{\overline{O},\overline{R}}(C)$.

Hence we get $\text{MAP-TD}^{\overline{O},R}(C) \le \text{MAP-TD}^{\overline{O},\overline{R}}(C)$ and $\text{MLE-TD}^{\overline{O},R}(C) \le \text{MLE-TD}^{\overline{O},\overline{R}}(C)$ despite the fact that $(\overline{O},R)$ does not dominate $(\overline{O},\overline{R})$.

5. On Concepts Taught by Labeled Examples

In this section, we will restrict ourselves to triples $(C,Z,\models)$ of the form as described in Example 2, i.e., C is a family of subsets of a domain X, $Z = X \times \{0,1\}$ and $\models$ is given by (2).

We will see that, for each triple $(C,Z,\models)$ of this special form and for each sampling mode (α,β) except (O,R), we have $\text{MAP-TD}^{\alpha,\beta}(C) = \mathrm{SMN}(G(C)^{\alpha,\beta})$. For $(\alpha,\beta) = (\overline{O},R)$, this is already known from Theorem 25. For the other sampling modes, $(\overline{O},\overline{R})$ and $(O,\overline{R})$, it will be shown in Section 5.1, Since the modes (O,R) and $(\overline{O},R)$ are equivalent, we see that, for triples of the special form, the MAP-teaching dimensions of C are fully determined by the saturating matching numbers associated with $G(C)$.

In Section 5.2 we explore how MAP- and MLE-learners are related. For a given collection of conditional likelihoods, it can make much of a difference whether we commit ourselves to uniform priors or not. However, in the case of optimally parameterized learners, the freedom for choosing a non-uniform prior is of minor importance only: it turns out that the MLE-teaching dimension exceeds the MAP-teaching dimension at most by 1.

In Section 5.3, we will see that the $\text{MLE-TD}^{\overline{O},\overline{R}}(C)$ is upper bounded by the so-called antichain number of C, by the VC-dimension of C and by the no-clash teaching dimension of C. These upper bounds are then, all the more, valid for all parameters $\text{MAP-TD}^{\alpha,\beta}(C)$ (no matter how the sampling mode (α,β)) is chosen).

In Section 5.4, we will show that the saturating matching numbers associated with $G(C)$ (and hence the MAP-teaching dimensions of C) can be computed in polytime.

5.1 Saturating Matching Number Revisited

We start with the two main results of this section.

Theorem 27 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2. Then* $\text{MAP-TD}^{\overline{O},\overline{R}}(C) = \text{SMN}(G(C)^{\overline{O},\overline{R}})$ *and* $\text{MLE-TD}^{\overline{O},\overline{R}}(C) = \text{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset})$.

Proof The $\geq$-direction of the claimed equalities is covered by Theorem 25. We have to show the $\leq$-direction. We may restrict ourselves to proving $\text{MLE-TD}^{\overline{O},\overline{R}}(C) \leq \text{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset})$ because the proof for $\text{MAP-TD}^{\overline{O},\overline{R}}(C) \leq \text{SMN}(G(C)^{\overline{O},\overline{R}})$ is quite similar and uses the same kind of arguments that we had used in the final part of the proof of Theorem 25.
Set $m = |X|$, $d^+ = \text{SMN}(G(C)^{\overline{O},\overline{R}})$ and let $M : C \rightarrow \mathcal{Z}^{\overline{O},\overline{R}} \setminus \{\emptyset\}$ be a C-saturating matching in $G(C)^{\overline{O},\overline{R}}$ of order d^+. For every $c \in C$, we set $d(c) = |M(c)|$. Note that $1 \leq d(c) \leq d^+$. If $d^+ = m$, then we are done because $\text{MLE-TD}^{\overline{O},\overline{R}}(C)$ cannot exceed m. We may assume therefore that $d^+ \leq m - 1$. Let $0 < \varepsilon \leq \frac{1}{2}$ be a small real number (where the meaning of "small" will become clear from what follows). For each $c \in C$, we set

$$U_0(c) := \{(x,b) \in Z : c(x) \neq b\} \quad , \quad U_1(c) := \{(x,b) \in Z : c(x) = b \wedge (x,b) \notin M(c)\} \tag{20}$$

and $U(c) = U_0(c) \cup U_1(c)$. Note that, for each $c \in C$, the set Z partitions into $M(c)$, $U_0(c)$ and $U_1(c)$. For each $c \in C$ and each $(x,b) \in Z$, we set

$$P((x,b)|c) = \begin{cases} \frac{1-\varepsilon}{d(c)} & \text{if } (x,b) \in M(c) \\ \frac{\varepsilon}{m-d(c)} & \text{if } (x,b) \in U_1(c) \\ 0 & \text{if } (x,b) \in U_0(c) \end{cases} \quad . \tag{21}$$

Let L be the MLE-learner given by (21). We aim at showing that the matching $M : C \rightarrow \mathcal{Z}^{\overline{O},\overline{R}} \setminus \{\emptyset\}$ can be seen as a teacher for L. To this end, it suffices to show that the condition

$$\forall c \neq c_0 \in C : P^{\overline{O},\overline{R}}(M(c_0)|c_0) > P^{\overline{O},\overline{R}}(M(c_0)|c) \tag{22}$$

is satisfied provided that ε is sufficiently small. We briefly note that $|M(c)| + |U_1(c)| = m \geq d^+$ and $\varepsilon \leq 1/2$, and proceed with two claims which will help us to verify (22).

Claim 1: Call a subset of Z *c-rare* if it contains a (low probability) element from $U(c)$ while missing a (high probability) element from $M(c)$. Suppose that $d \leq d^+$. Then the probability that d-fold $P(\cdot|c)$-sampling without replacement leads to a c-rare sample is smaller than $d\varepsilon$ divided by $\frac{1-\varepsilon}{d(c)}$ and, therefore, smaller than $2dd(c)\varepsilon$.

Proof of Claim 1: The total $P(\cdot|c)$ probability mass of $U(c)$ is ε whereas any element of $M(c)$ has a $P(\cdot, c)$-probability of $\frac{1-\varepsilon}{d(c)}$. For $k = 1, \ldots, d$, let E_k be the event that, within trial k, a point from $U(c)$ is sampled although at least one point from $M(c)$ has not been sampled before. It suffices to upper-bound the probability of $E_1 \vee \ldots \vee E_d$. The probability of E_k is obviously smaller than ε divided by $\frac{1-\varepsilon}{d(c)}$ and therefore smaller than $\frac{d(c)\varepsilon}{1-\varepsilon} \leq 2d(c)\varepsilon$. An application of the union bound yields an additional factor d.

Claim 2: Suppose that $d \leq d(c)$. Then a sample of size d which contains an element from $U_1(c)$ is c-rare (because it necessarily must miss an element from $M(c)$).

Setting $c = c_0$ and $d = d(c_0)$, we infer from the above claims that $P^{\overline{O},\overline{R}}(M(c_0)|c_0) > 1 - 2d(c_0)^2\varepsilon$. Consider now an arbitrary, but fixed, concept $c_1 \in C\setminus\{c_0\}$. Then $M(c_1) \neq M(c_0)$. We proceed by case analysis:

Case 1: Neither $M(c_0) \subset M(c_1)$ nor $M(c_1) \subset M(c_0)$.
Then $M(c_0)$ is a c_1-rare sample. Hence $P^{\overline{O},\overline{R}}(M(c_0)|c_1) < 2d(c_0)d(c_1)\varepsilon$.

Case 2: $M(c_0) \subset M(c_1)$.
We apply a symmetry argument. Every sample containing $d(c_0)$ elements of $M(c_1)$ has the same chance for being obtained from $d(c_0)$-fold $P(\cdot|c_1)$-sampling without replacement. Hence

$$P^{\overline{O},\overline{R}}(M(c_0)|c_1) \leq \binom{d(c_1)}{d(c_0)}^{-1} \leq \frac{1}{d(c_1)} \leq \frac{1}{2} \ ,$$

where the last two inequalities follow from $1 \leq d(c_0) \leq d(c_1) - 1$.

Case 3: $M(c_1) \subset M(c_0)$.
We may assume that $M(c_0) \subseteq M(c_1) \cup U_1(c_1)$ because, otherwise, we obtain directly $P^{\overline{O},\overline{R}}(M(c_0)|c_1) = 0$. We apply again a symmetry argument. Every sample containing $M(c_1)$ and $d(c_0) - d(c_1)$ elements of $U_1(c_1)$ has the same chance for being obtained from $d(c_0)$-fold $P(\cdot|c_1)$-sampling without replacement. Hence

$$P^{\overline{O},\overline{R}}(M(c_0)|c_1) \leq \binom{m - d(c_1)}{d(c_0) - d(c_1)}^{-1} \ .$$

The latter expression is upper-bounded by $\frac{1}{2}$ because $1 \leq d(c_0) - d(c_1) < m - d(c_1)$, $d(c_1) \leq d(c_0) - 1 \leq m - 2$ and, therefore, $m - d(c_1) \geq 2$.

It becomes obvious from this discussion that condition (22) is satisfied provided that ε is sufficiently small. ∎

Theorem 28 *Suppose that* $(C, Z, \models)$ *is of the form as described in Example 2. Then* $\text{MAP-TD}^{O,\overline{R}}(C) = \text{SMN}(G(C)^{O,\overline{R}})$ *and* $\text{MLE-TD}^{O,\overline{R}}(C) = \text{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})$.

Proof The $\geq$-direction of the claimed equalities is covered by Theorem 25. We have to show the $\leq$-direction. We may restrict ourselves to proving $\text{MLE-TD}^{O,\overline{R}}(C) \leq \text{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})$ because the proof for $\text{MAP-TD}^{O,\overline{R}}(C) \leq \text{SMN}(G(C)^{O,\overline{R}})$ is quite similar and uses the same kind of arguments that we had used in the final part of the proof of Theorem 25.
Set $m = |X|$, $d^+ = \text{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})$ and let $M : C \to \mathcal{Z}^{O,\overline{R}}\setminus\{\emptyset\}$ be a C-saturating matching in $G(C)^{O,\overline{R}}_{\neq\emptyset}$ of order d^+. If $d^+ = m$, then we are done because $\text{MLE-TD}^{O,\overline{R}}(C)$ cannot exceed m. We may assume therefore that $d^+ \leq m-1$. For every $c \in C$, we set $d(c) = |M(c)|$. Note that $1 \leq d(c) \leq d^+$. We fix for each concept $c \in C$ a sequence $z_1^c, \ldots, z_m^c$ consisting of all elements of Z_c subject to the constraint that $z_1^c, \ldots, z_{d(c)}^c = M(c)$, i.e., this sequence must start with $M(c)$. In the sequel, we will specify the parameter set of an MLE-learner

of C. We do this in two stages. In Stage 1, we make a preliminary definition which already achieves that each $c^* \in C$ is a (not necessarily unique) maximizer of $P^{O,\overline{R}}(M(c^*|c))$. In Stage 2, we make some infinitesimal changes of the parameter set (by bringing a small parameter $\varepsilon > 0$ into play) so that, after these changes have taken place, each $c^* \in C$ will be a unique maximizer of $P^{O,\overline{R}}(M(c^*|c))$. This would imply that M can be viewed as a teacher for L, which would complete the proof. Details follow.

We enter Stage 1 of the parameter construction. Let L be the MLE-learner whose parameters are given by

$$P(z|c) = \begin{cases} 2^{-i} & \text{if } 1 \le i \le d(c) \text{ and } z = z_i^c \\ \frac{2^{-d(c)}}{m-d(c)} & \text{if } d(c)+1 \le i \le m \text{ and } z = z_i^c \\ 0 & \text{if } z \in Z \setminus Z_c \end{cases} \quad .$$

In other words, given c, L assigns probability mass 2^{-i} to the i-the element of the sequence $M(c)$ and distributes the remaining probability mass, $2^{-d(c)}$, evenly on the elements of $Z_c \setminus M(c)$. Note that the c-conditional likelihood of an element in $M(c)$ is at least $2^{-d(c)}$ while the probability of an element in $Z_c \setminus M(c)$ equals $\frac{2^{-d(c)}}{m-d(c)} \le 2^{-d(c)}$ with equality only if $d(c) = m-1$. It is easy to determine the c-conditional likelihood of $M(c)$:

$$P^{O,\overline{R}}(M(c)|c) = \frac{\prod_{i=1}^{d(c)} 2^{-i}}{\prod_{i=1}^{d(c)-1} 2^{-i}} = 2^{-d(c)} \quad .$$

The middle term contains in the numerator the product of the c-conditional likelihoods of $z_1^c, \ldots, z_{d(c)}^c$, respectively. In the denominator, it contains the product of the corresponding normalization factors: if $z_1^c, \ldots, z_j^c$ haven been sampled within the first j trials, then the remaining probability mass equals $1 - \sum_{i=1}^j 2^{-i} = 2^{-j}$. Let us now fix an arbitrary target concept $c^* \in C$ and see how the c^*-conditional likelihood of $M(c^*)$ relates to the c-conditional likelihood of $M(c^*)$ for some other concept $c \in C \setminus \{c^*\}$. We aim at showing that $P^{O,\overline{R}}(M(c^*)|c) \le P^{O,\overline{R}}(M(c^*)|c^*)$. We may assume that $c \models M(c^*)$ because, otherwise, we would obtain $P^{O,\overline{R}}(M(c^*)|c) = 0$, and we were done. For sake of simplicity, we set $d := d(c^*)$ and $z_i := z_i^{c^*}$ for $i = 1, \ldots, d$.

Let us briefly discuss the case that $M(c)$ and $M(c^*)$ are equal as sets. Then there exists a permutation σ such that $M(c) = z_{\sigma(1)}, \ldots, z_{\sigma(d)}$. Since M is a matching, σ cannot be the identity permutation. It follows that $P^{O,\overline{R}}(M(c^*)|c^*) > P^{O,\overline{R}}(M(c^*)|c)$ because $(P(z_i|c^*))_{i=1,\ldots,d} = (2^{-i})_{i=1,\ldots,d}$ is a strictly decreasing sequence while $(P(z_i|c))_{i=1,\ldots,d}$ (as a non-identity permutation of $(2^{-i})_{i=1,\ldots,d}$) is not.[11]

From now, we assume that $M(c)$ and $M(c^*)$ are different even when viewed as sets. Let j be the number of $z \in Z$ occurring in $M(c)$ and in $M(c^*)$. We can make the pessimistic assumption that the sequence $M(c)$ starts with $z_1, \ldots, z_j$ because this will lead to the largest conceivable value of $P^{O,\overline{R}}(M(c^*)|c)$.[12] The remaining observations $z_{j+1}, \ldots, z_{d(c)}$ must then be members of $Z_c \setminus M(c)$. Remember that for each $z \in Z_c \setminus M(c)$ we have that $P(z|c) = \frac{2^{-d(c)}}{m-d(c)}$. The term $P^{O,\overline{R}}(M(c^*)|c)$ can be expressed as a product of two terms.

[11] Compare with Remark 4.

[12] This brings the j largest c-conditional likelihoods into play and puts them in the most effective position.

The first one (resp. second one) is the contribution of the first j trials (resp. the last $d-j$ trials). Since $M(c)$ starts with $z_1, \ldots, z_j$, the first term is simply $T_1 := 2^{-j}$. The second term has the following form

$$T_2 := \frac{\left(\frac{2^{-j}}{m-j}\right)^{d-j}}{2^{-j}\left(2^{-j} - \frac{2^{-j}}{m-j}\right)\left(2^{-j} - 2\frac{2^{-j}}{m-j}\right)\ldots\left(2^{-j} - (d-j-1)\frac{2^{-j}}{m-j}\right)} \enspace .$$

As usual, the numerator contains the product of the c-conditional (here: uniform) likelihoods while the denominator contains the product of the corresponding normalization factors. T_2 looks terrifying at first glance, but luckily there is a lot of cancellation and T_2 can be rewritten as follows:

$$\begin{aligned} T_2 &= \frac{1}{(m-j)^{d-j}\left(1 - \frac{1}{m-j}\right)\left(1 - \frac{2}{m-j}\right)\ldots\left(1 - \frac{d-j-1}{m-j}\right)} \\ &= \frac{1}{(m-j)(m-j-1)(m-j-2)\ldots(m-d+1)} \enspace . \end{aligned}$$

Remember that $d = d(c^*) \le m-1$. It follows that $m-d+1 \ge 2$ and therefore

$$T_2 \le 2^{-(d-j)} \text{ and } P^{O,\overline{R}}(M(c^*)|c) = T_1 \cdot T_2 \le 2^{-d}$$

with equality only if either $j = d$ or $d = m-1$ and $j = m-2$. Note that $j = d$ if and only if the sequence $M(c)$ starts with the sequence $M(c^*) = z_1, \ldots, z_d$.

We enter now Stage 2 of the parameter construction, in which we make some infinitesimal changes of the parameters that we have used so far. In order to distinguish the new parameter collection from the old one, the new parameters are denoted by $P_\varepsilon(z|c)$. They are defined as follows:

$$P_\varepsilon(z|c) = \begin{cases} 2^{-i} & \text{if } 1 \le i \le d(c)-1 \text{ and } z = z_i^c \\ 2^{-i} + \varepsilon & \text{if } i = d(c) \text{ and } z = z_i^c \\ \frac{2^{-d(c)} - \varepsilon}{m-d(c)} & \text{if } d(c)+1 \le i \le m \text{ and } z = z_i^c \\ 0 & \text{if } z \in Z \setminus Z_c \end{cases} \enspace .$$

The main difference to the old parameter collection is the "extra-bonus" ε that c assigns to the last element $z_{d(c)}^c$ of the sequence $M(c)$. Now the total probability mass assigned to $z_1^c, \ldots, z_{d(c)}^c$ is by the amount of ε greater than before, so that only probability mass $2^{-d(c)} - \varepsilon$ is left for $Z_c \setminus M(c)$. Again, this probability mass is shared evenly among the elements of $Z_c \setminus M(c)$. Here comes the central observation:

Claim: If $\varepsilon > 0$ is sufficiently small, then the following implications are valid:

$$\begin{aligned} P^{O,\overline{R}}(M(c^*)|c^*) > P^{O,\overline{R}}(M(c^*)|c) &\Rightarrow P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) > P_\varepsilon^{O,\overline{R}}(M(c^*)|c) \enspace , \\ P^{O,\overline{R}}(M(c^*)|c^*) = P^{O,\overline{R}}(M(c^*)|c) &\Rightarrow P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) > P_\varepsilon^{O,\overline{R}}(M(c^*)|c) \enspace . \end{aligned}$$

Proof of the Claim: The first implication is based on a simple continuity argument. The second implication can be verified as follows. Remember from the discussion in Stage

1 that $P^{O,\overline{R}}(M(c^*)|c^*) = P^{O,\overline{R}}(M(c^*)|c)$ can occur only if either $M(c)$ starts with $M(c^*) = z_1, \ldots, z_d$ or if $d = m-1$ and $j = m-2$. In the former case, the effect of $P_\varepsilon(z_d|c^*) = P(z_d|c^*) + \varepsilon$ and $P_\varepsilon(z_d|c) = P(z_d|c)$ will be that

$$P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) > P^{O,\overline{R}}(M(c^*)|c^*) = P^{O,\overline{R}}(M(c^*)|c) = P_\varepsilon^{O,\overline{R}}(M(c^*)|c) \ ,$$

as desired. In the latter case, we have $M(c^*) = z_1, \ldots, z_{m-1}$ and either $M(c) = z_1, \ldots, z_{m-2}$ or $M(c) = z_1, \ldots, z_{m-2}, z_m$. In the latter case, we obtain

$$P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) > P^{O,\overline{R}}(M(c^*)|c^*) = P^{O,\overline{R}}(M(c^*)|c) > P_\varepsilon^{O,\overline{R}}(M(c^*)|c) \ ,$$

which is again the desired result. Suppose therefore that $M(c^*) = z_1, \ldots, z_{m-1}$ and $M(c) = z_1, \ldots, z_{m-2}$. Here the situation is less clear, because the ε-bonus will affect not only the c^*-conditional likelihood of $M(c^*)$ but also the c-conditional likelihood. We therefore compute both quantities and compare them afterwards. Clearly $P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) = 2^{-(m-1)} + \varepsilon$. The term $P_\varepsilon^{O,\overline{R}}(M(c^*)|c)$ can be expressed as a product of two terms, The first one (resp. second one) is the contribution of the first $m-2$ trials (resp. the last trial). Since $M(c) = z_1, \ldots, z_{m-2}$, the first term clearly equals $2^{-(m-2)} + \varepsilon$. Note that $2^{-(m-2)} - \varepsilon$ is the probability mass remaining for, and evenly shared by, z_{m-1} and z_m. The second term equals therefore

$$\frac{P_\varepsilon(z_{m-1}|c)}{2^{-(m-2)} - \varepsilon} = \frac{\left(2^{-(m-2)} - \varepsilon\right)/2}{2^{-(m-2)} - \varepsilon} = \frac{1}{2} \ .$$

It follows that

$$P_\varepsilon^{O,\overline{R}}(M(c^*)|c) = \frac{1}{2} \cdot \left(2^{-(m-2)} + \varepsilon\right) = 2^{-(m-1)} + \frac{\varepsilon}{2} \ ,$$

which is less than $P_\varepsilon^{O,\overline{R}}(M(c^*)|c^*) = 2^{-(m-1)} + \varepsilon$. This completes the proof of the claim.

The above discussions show that we can view M a teacher for the learner L with parameter collection $(P_\varepsilon(z|c))_{z \in Z, c \in C}$. This completes the proof of the theorem. ■

Combining Theorems 27 and 28 with what we already know about saturating matching numbers, we obtain the following result:

Corollary 29 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2 and $(\alpha, \beta) \neq (O, R)$. Then*

$$\text{MAP-TD}^{\alpha,\beta}(C) = \text{SMN}(G(C)^{\alpha,\beta}) \quad \textit{and} \quad \text{MLE-TD}^{\alpha,\beta}(C) = \text{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset}) \ .$$

Moreover

$$\begin{aligned} \text{MAP-TD}^{\overline{O},\overline{R}}(C) &\geq \max\{\text{MAP-TD}^{O,R}(C), \text{MAP-TD}^{O,\overline{R}}(C)\} \ , \\ \text{MLE-TD}^{\overline{O},\overline{R}}(C) &\geq \max\{\text{MLE-TD}^{O,R}(C), \text{MLE-TD}^{O,\overline{R}}(C)\} \ . \end{aligned}$$

The first assertion of the corollary implies the correctness of the results which are visualized in Fig. 1. The following result provides some supplementary information:

Theorem 30 *Let (α,β) and (α',β') be two different sampling modes. There exists a concept class C such that* $\mathrm{SMN}(G(C)^{\alpha',\beta'}) \neq \mathrm{SMN}(G(C)^{\alpha,\beta})$.

Proof We present the proof for $(\alpha,\beta) = (\overline{O},R)$ and $(\alpha',\beta') = (\overline{O},\overline{R})$.[13] Let $X = \{x_1,\ldots,x_m\}$, let $Z = X \times \{0,1\}$, let C_m be the powerset of X and let $\models$ be given by (2). Let $\mathcal{Z}_2$ (resp. $\mathcal{Z}'_2$) be the set of all $A \in \mathcal{Z}^{\overline{O},R}$ (resp. $A \in \mathcal{Z}^{\overline{O},\overline{R}}$) such that $|A| \le 2$. A simple counting argument shows that $|\mathcal{Z}'_2| < |\mathcal{Z}_2|$. Consider the bipartite graph G with vertex sets C_m and $\mathcal{Z}_2$ and with an edge (c,A) if and only if $c \models A$. Each vertex in $\mathcal{Z}_2$ has degree at least $D := 2^{m-2}$ whereas each vertex in C_m has degree $d := 1 + 2m + \frac{1}{2}(m-1)m$. Suppose that m is sufficiently large such that $d \le D$. Fix an arbitrary subset S of $\mathcal{Z}_2$. It follows that

$$|\Gamma(S)| \ge \frac{D}{d} \cdot |S| \ge |S|$$

so that G satisfies Hall's condition. It follows that G admits a $\mathcal{Z}_2$-saturating matching, say M. Let C be the set of concepts in C_m having an M-partner. By construction: $\mathrm{SMN}(G(C)^{\overline{O},R}) = 2$. For cardinality reasons, namely $|C| = |M| = |\mathcal{Z}_2| > |\mathcal{Z}'_2|$, we have $\mathrm{SMN}(G(C)^{\overline{O},\overline{R}}) > 2$. ■

Theorem 30 implies that the parameters with different colors in Fig. 1 can generally have different values.

5.2 MAP- versus MLE-Learners

Suppose that L is an MLE-learner for C. Let L' be a MAP-learner that differs from L only by having non-uniform priors, i.e., the conditional likelihoods are the same. The following example demonstrates that the gap between $\text{MAP-TD}_L^{\alpha,\beta}(C)$ and $\text{MAP-TD}_{L'}^{\alpha,\beta}(C)$ can become arbitrarily large.[14]

Example 4 *Let $X = \{x_1,\ldots,x_m\}$, $Z = X \times \{0,1\}$, $C = \{\{x_1\},\ldots,\{x_m\}\} \cup \{\emptyset\}$ and let $\models$ be given by (2). Consider the MLE-learner L be given by the parameters*

$$P((x_i,c(x_i))|c) \;=\; \frac{1}{m}$$

for each $c \in C$ and $i = 1,\ldots,m$. We assume for simplicity that the sampling mode (α,β) of L equals $(\overline{O},\overline{R})$, but the following reasoning (mutatis mutandis) applies to any other sampling mode as well. Clearly, for each $k \in [m]$, the concept $\{x_k\}$ can be taught by the single observation $(x_k,1)$. However $\emptyset$ can only be taught by the full set $A_0 := \{(x_i,0) : i = 1,\ldots,m\}$ of observations that $\emptyset$ is consistent with: as long as some $(x_k,0)$ is missing in a set $A \subset A_0$, we have that $P(A|\emptyset) = P(A|\{x_k\})$ so that $\emptyset$ is not the unique maximizer

[13] The proof for the other choices of (α,β) and (α',β') is similar.

[14] This example uses a concept class, namely singletons plus empty set, which is often used to demonstrate that the classical teaching model from Shinohara and Miyano (1991); Goldman and Kearns (1995) may assign an inappropriately high teaching dimension to a trivial concept class.

of $P(A|c)$. We may conclude from this discussion that $\text{MAP-TD}_L^{\alpha,\beta}(C) = m$. *Let L' be a MAP-learner that differs from L only by having for $\emptyset$ a higher prior than for the other concepts in C. Then the concept $\{x_k\}$ can still be taught by the single observation $(x_k, 1)$. But now also the concept $\emptyset \in C$ can be taught in a trivial fashion by $\emptyset \in 2^Z$. We may conclude that* $\text{MAP-TD}_{L'}^{\alpha,\beta}(C) = 1$.

In contrast to Example 4, the next result shows that, in case of optimally parameterized learners, the advantage of MAP-learners over MLE-learners is anything but dramatic:

Theorem 31 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2 and $(\alpha, \beta) \neq (O, R)$. Then*

$$\text{MAP-TD}^{\alpha,\beta}(C) \leq \text{MLE-TD}^{\alpha,\beta}(C) \leq 1 + \text{MAP-TD}^{\alpha,\beta}(C) \ . \tag{23}$$

Moreover, there exist concept classes C' and C'' such that

$$\text{MLE-TD}^{\alpha,\beta}(C') = \text{MAP-TD}^{\alpha,\beta}(C') \quad \textit{and} \quad \text{MLE-TD}^{\alpha,\beta}(C'') = 1 + \text{MAP-TD}^{\alpha,\beta}(C'') \ . \tag{24}$$

Proof Clearly $\text{MAP-TD}^{\alpha,\beta}(C) \leq \text{MLE-TD}^{\alpha,\beta}(C)$. In order to obtain (23), it suffices therefore to show that $\text{MLE-TD}^{\alpha,\beta}(C) \leq 1+\text{MAP-TD}^{\alpha,\beta}(C)$, or equivalently, that $\text{SMN}(G(C)_{\neq\emptyset}^{\alpha,\beta}) \leq 1 + \text{SMN}(G(C)^{\alpha,\beta})$. We present the proof for $(\alpha, \beta) = (\overline{O}, \overline{R})$.[15] For sake of brevity, set $m := |X|$, $G = G(C)^{\overline{O},\overline{R}}$ and $d := \text{SMN}(G)$. Since $\text{SMN}(G_{\neq\emptyset}) \leq m$, we may assume that $d \leq m - 2$. Let $M : C \to 2^Z$ be a C-saturating matching of order d in G. If M does not assign $\emptyset$ to any concept in C, then $\text{SMN}(G_{\neq\emptyset}) \leq d$. Otherwise, if $M(c_0) = \emptyset$ for some $c_0 \in C$, then we may arbitrarily pick a set $A \subset X$ of size $d + 1$ and replace the M-partner $\emptyset$ of c_0 by the set $B = \{(a, c_0(a)) : a \in A\}$. The resulting matching now witnesses that $\text{SMN}(G_{\neq\emptyset}) \leq d + 1$.
We still have to specify concept classes C' and C'' which satisfy (24). As for C', there are plenty of choices, e.g., $C' = \{\{x_i\} : i = 1, \ldots, m\}$ satisfies

$$\text{MLE-TD}^{\alpha,\beta}(C') = \text{MAP-TD}^{\alpha,\beta}(C') = 1 \ .$$

In order to specify an appropriate class C'', we assume again that $(\alpha, \beta) = (\overline{O}, \overline{R})$ and proceed as follows. Let $X = \{x_1, \ldots, x_m\}$, let $Z = X \times \{0, 1\}$, let C_m be the powerset of X and let $\models$ be given by (2). Let $\mathcal{Z}_{\leq d}$ (resp. $\mathcal{Z}'_{\leq d}$) be the set of subsets (resp. non-empty subsets) of Z of size at most d. Consider the bipartite graph G with vertex sets C_m and $\mathcal{Z}_{\leq d}$ an edge (c, A) if and only if $c \models A$. If m is sufficiently large (while d is kept fixed), G admits a $\mathcal{Z}_{\leq d}$-saturating matching, say M. Let C'' be the set of concepts in C_m having an M-partner. By construction: $\text{SMN}(G(C'')^{\overline{O},\overline{R}}) = d$. For cardinality reasons, namely $|C''| = |M| = |\mathcal{Z}_{\leq d}| > |\mathcal{Z}_{\leq d}| - 1 = |\mathcal{Z}'_{\leq d}|$, we have $\text{SMN}(G(C'')_{\neq\emptyset}^{\overline{O},\overline{R}}) > d$, which implies that $\text{SMN}(G(C'')_{\neq\emptyset}^{\overline{O},\overline{R}}) = d + 1$. ■

[15] The proof for the other choices of (α, β) is similar.

5.3 Parameters Bounding MLE-TD from Above

Since MLE-TD can never be smaller than MAP-TD, it follows that MLE-TD$^{\overline{O},\overline{R}}(C)$ is the largest among the parameters occurring in Corollary 29. Hence upper bounds on MLE-TD$^{\overline{O},\overline{R}}(C)$ are, all the more, upper bounds on the other parameters. For this reason, we confine ourselves to MLE-learners and to sampling mode $(\overline{O},\overline{R})$ in what follows. In order to simplify notation, we will write

- 2^Z instead of $\mathcal{Z}^{\overline{O},\overline{R}}$,
- MLE-TD(C) instead of MLE-TD$^{\overline{O},\overline{R}}(C)$,
- $G^+(C)$ instead of $G(C)^{\overline{O},\overline{R}}_{\neq\emptyset}$.

Among the parameters that bound MLE-TD(C) from above are the antichain number of C, the VC-dimension of C and the so-called no-clash teaching dimension of C. We begin with the definition of the antichain number:

Definition 32 (Antichain Mapping and Antichain Number) *$T : C \rightarrow 2^Z$ is called an* antichain mapping for C *if the following holds:*

1. *Each concept $c \in C$ is consistent with $T(c)$.*
2. *The sets $(T(c))_{c\in C}$ form an antichain, i.e.,*

$$\forall c_1 \neq c_2 \in C : T(c_1) \not\subseteq T(c_2) \wedge T(c_2) \not\subseteq T(c_1) \ \ .$$

The smallest possible order of an antichain mapping for C is called the antichain number of C *and denoted by* AN(C).

It is well-known that the antichain number is upper-bounded by the VC-dimension:

Theorem 33 (Mansouri et al. (2022)) *Suppose that the concept class C is a family of subsets of a finite domain X. Then* AN$(C) \leq$ VCdim(C).

We proceed with the definition of the teaching dimension in the so-called no-clash model of teaching:

Definition 34 (No-clash Teaching Dimension Kirkpatrick et al. (2019); Fallat et al. (2022)) *A mapping $T : C \rightarrow 2^Z$ is called* clash-free on C *if it satisfies the following:*

1. *Each $c \in C$ is consistent with $T(c)$.*
2. *If $c_1 \neq c_2 \in C$, then c_1 is inconsistent with $T(c_2)$ or c_2 is inconsistent with $T(c_1)$.*[16]

The no-clash teaching dimension of C, *denoted as* NC-TD(C), *is the smallest possible order of a mapping $T : C \rightarrow 2^Z$ that is clash-free on C.*

[16] The situation that c_1 is consistent with $T(c_2)$ and c_2 is consistent with $T(c_1)$ would be called a *clash of c_1 and c_2*. This explains why the mapping T is called clash-free.

Theorem 35 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2. Then* $\text{MLE-TD}(C) \le \text{AN}(C)$ *and* $\text{MLE-TD}(C) \le \text{NC-TD}(C)$.

Proof Because $\text{MLE-TD}(C) = \text{SMN}(G^+(C))$, it suffices to show that $\text{SMN}(G^+(C))$ is upper-bounded by $\text{AN}(C)$ and $\text{NC-TD}(C)$. An antichain mapping $T : C \to 2^Z$ clearly satisfies (17) and does not have $\emptyset$ in its image. Thus, an application of Remark 24 yields $\text{AN}(C) \ge \text{SMN}(G^+(C))$. A clash-free mapping $T : C \to 2^Z$ must be of order at least 1. There can be at most one concept c in C such that $T(c) = \emptyset$. Suppose that $T(c) = \emptyset$. Consider an arbitrary, but fixed, concept $c' \in C \setminus \{c\}$. Since c' is consistent with (the empty sample) $T(c)$ and T is clash-free, the concept c must be inconsistent with $T(c')$. Let us redefine $T(c)$ as a singleton set $\{(x, b)\}$ such that $b = c(x)$. This modification of T is still clash-free and leaves the order of T unchanged. Moreover, after this modification, T satisfies (17) and does not have $\emptyset$ in its image. Now another application of Remark 24 yields $\text{NC-TD}(C) \ge \text{SMN}(G^+(C))$. ■

The inequality $\text{MLE-TD}(C) \le \text{NC-TD}(C)$ had been proven already in Ferri et al. (2022). The proof given there does not make use of saturating matching numbers and is more complicated. Because $\text{AN}(C) \le \text{VCdim}(C)$, we immediately obtain the following result:

Corollary 36 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2. Then* $\text{MLE-TD}(C) \le \text{VCdim}(C)$.

5.4 Computational Considerations

We will show in the course of this section that $\text{SMN}(G^+(C))$ (and related quantities) can be computed in time $\text{poly}(|C|, |X|)$ from a given (finite) concept class $C \subseteq 2^X$. The central observation will be that, in order to find a C-saturating matching of minimum order in $G^+(C)$, we do not need to compute the (possibly exponentially large) bipartite graph $G^+(C)$. All pieces of information about $G^+(C)$ that we need in the course of the algorithm can be efficiently extracted from the much smaller bipartite graph $G(C)$.

We start with a lemma that is particularly interesting when we have a bipartite graph whose first vertex set, V_1, is much smaller than its second vertex set, V_2:

Lemma 37 *Let $G = (V_1, V_2, E)$ with $E \subseteq V_1 \times V_2$ be a bipartite graph. Let $\mathcal{O}$ be an oracle that, upon request (v, k) with $v \in V_1$ and $k \in [|V_1|]$, returns* $\min\{\deg_G(v), k\}$ *distinct neighbors of v.*[17] *Then there is an oracle algorithm $A^{\mathcal{O}}$ which computes a maximum matching in G and has a time bound that is polynomial in $|V_1|$.*

Proof For sake of brevity, we set $n = |V_1|$. Let $V_1' \subseteq V_1$ be the set of vertices in V_1 with less than n neighbors, and let $V_1'' = V_1 \setminus V_1'$ be the set of remaining vertices in V_1, i.e., the vertices with at least n neighbors. The algorithm $A^{\mathcal{O}}$ proceeds as follows:

1. For each $v \in V_1$, it sends the request (v, n) to $\mathcal{O}$ and receives a list of all neighbors if $v \in V_1'$, resp. a list of n distinct neighbors if $v \in V_1''$.

[17] The oracle $\mathcal{O}$ can be implemented efficiently if, for instance, G is represented by the adjacency lists for the vertices in V_1 and there is direct access to each of these lists.

2. Let G' be the subgraph of G that is induced by V_1' and $\Gamma(V_1')$. Let MAX-MATCH be a standard polynomial-time algorithm for maximum matching computation. $A^{\mathcal{O}}$ applies MAX-MATCH to G' and MAX-MATCH returns a maximum matching M' in G'.

3. $A^{\mathcal{O}}$ augments M' to a V_1-saturating matching in a greedy fashion: for each $v \in V_1''$, it inspects the list of n distinct neighbors of v and matches v with the first neighbor which had not been matched before.

Note that G' has at most $n(n-1)$ vertices. Moreover, among n neighbors of a vertex $v \in V_1''$, there must be at least 1 neighbor which is not already matched with another vertex in V_1. It easily follows that $A^{\mathcal{O}}$ returns a maximum matching in $\mathrm{poly}(|V_1|)$ time. ■

With a bipartite graph $G = (V_1, V_2, E)$, we associate the bipartite graph

$$G^+ = (V_1, 2^{V_2} \setminus \{\emptyset\}, E^+) \;\text{ with }\; E^+ = \{(v, B) \in V_1 \times 2^{V_2} \setminus \{\emptyset\} : \{v\} \times B \subseteq E\}\ . \tag{25}$$

In other words: the pair (v, B) with $v \in V_1$ and $\emptyset \subset B \subseteq V_2$ is an edge in E^+ iff, for every $v' \in B$, the pair (v, v') is an edge in E.

Theorem 38 *Given a bipartite graph $G = (V_1, V_2, E)$, a V_1-saturating matching of minimum order in G^+ (resp. an error message if a V_1-saturating matching does not exist) can be computed in polynomial time.*

Proof We consider first the problem of computing a V_1-saturating matching of minimum order in G^+. Let us fix some notation. For $\ell = 1, \ldots, |V_2|$, let $G^{(\ell)} = (V_1, V_2^{(\ell)}, E^{(\ell)})$ be the bipartite graph given by

$$V_2^{(\ell)} = \{B \subseteq V_2 : 1 \le |B| \le \ell\}) \;\text{ and }\; E_2^{(\ell)} = \{(v, B) \in V_1 \times V_2^{(\ell)} : \{v\} \times B \subseteq E\}\ .$$

In other words, $G^{(\ell)}$ is the subgraph of G^+ induced by V_1 and $V_2^{(\ell)}$. Given G, $\ell \in [|V_2|]$, $k \in [|V_1|]$ and $v \in V_1$, it is easy to compute a list of $\min\{\deg(v), k\}$ distinct neighbors of v in $G^{(\ell)}$. It follows from Lemma 37 that, given G and $\ell \in [|V_2|]$, we can compute in $\mathrm{poly}(|V_1|, |V_2|)$ steps a maximum matching M_ℓ in $G^{(\ell)}$. Let ℓ^+ be the minimum ℓ such that M_ℓ is of size $|V_1|$, respectively $\ell^+ = 1 + |V_2|$ if none of the M_ℓ saturates V_1. If $\ell^+ \le |V_2|$, then M_{ℓ^+} is the desired V_1-saturating matching of minimum order in G^+. If $\ell^+ = |V_2| + 1$, we may report error because G^+ does not admit a V_1-saturating matching. ■

Corollary 39 *Suppose that $(C, Z, \models)$ is of the form as described in Example 2. Then the following objects can be computed in polynomial time:*

- *the bipartite consistency graph $G(C)$ with vertex sets C and Z*
- *the (identical) parameters* $\mathrm{SMN}(G^+(C))$ *and* $\text{MLE-TD}(C)$
- *a C-saturating matching M in $G^+(C)$ of order* $\mathrm{SMN}(G^+(C))$

- *parameters representing an MLE-learner L for C and a teacher T for L who is of order* MLE-TD(C)

Proof Given C, the set Z and the bipartite graph $G(C)$ can clearly be computed in polynomial time. We may now apply Theorem 38 to the bipartite graph $G = G(C)$. Then G^+ in Theorem 38 equals $G^+(C)$. Hence the algorithm sketched in the proof of Theorem 38 can be used for finding a C-saturating matching M in $G^+(C)$ of minimum order (which is order $\mathrm{SMN}(G^+(C))$). As a byproduct, the parameter $\mathrm{SMN}(G^+(C))$ is now known as well. As for the specification of an appropriate MLE-learner L, we may use the parameter setting that is found in the proof of Theorem 25. As also shown in that proof, M (already known to be computable from C in polynomial time) represents a teacher of order MLE-TD(C) for L. This completes the proof of the corollary. ■

It is straightforward to extend Corollary 39 from sampling mode $(\overline{O}, \overline{R})$ to other sampling modes, and from MLE-TD to MAP-TD. The main point is to adjust the definition of G^+ in (25) so that $G(C)^+$ becomes identical to $G(C)^{\alpha,\beta}_{\neq\emptyset}$ resp. to $G(C)^{\alpha,\beta}$. We omit the details.

Open Problems and Future Work. What are “natural parameterizations” of MAP- or MLE-learners? Does MAP-based teaching of naturally parameterized learners force the teacher to present observations/examples which illustrate the underlying target concept in an intuitively appealing way?

Acknowledgements We would like to thank the reviewers for their very useful comments and suggestions.

Appendix A. Proof of Facts 1–4

Fact 1: Suppose that $0 \le |\Delta| < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then $P^{O,R}(z_1, z_2)|c^{\pm\Delta})$ and $P^{\overline{O},\overline{R}}(z_1, z_2|c^{\pm\Delta})$ are both strictly decreasing when $|\Delta|$ is increased.

Proof The assertion is obvious for $P^{O,R}(z_1, z_2)|c^{\pm\Delta}) = (p+\Delta)(p-\Delta) = p^2-\Delta^2$. Consider now the function

$$h(\Delta) := P^{\overline{O},\overline{R}}(z_1, z_2|c^{\pm\Delta}) = \frac{(p+\Delta)(p-\Delta)}{1-p-\Delta} + \frac{(p-\Delta)(p+\Delta)}{1-p+\Delta} = \frac{2(1-p)(p^2-\Delta^2)}{(1-p)^2-\Delta^2} \ ,$$

where the last equation can be obtained by a straightforward calculation. Another straightforward, but tedious, calculation shows that

$$h'(\Delta) = -\frac{4(1-p)(1-2p)\Delta}{((1-p)^2-\Delta^2)^2} .$$

Hence the function $h(\Delta)$ is strictly increasing for $\Delta < 0$ and strictly decreasing for $\Delta > 0$. It is therefore strictly decreasing when $|\Delta|$ is increased. ■

Fact 2: Suppose that $0 \le \Delta < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then

$$P^{O,\overline{R}}(z_1, z_2|c^{\pm\Delta}) - P^{O,\overline{R}}(z_1, z_2|c^{\pm 0}) \begin{cases} = 0 & \text{if } \Delta \in \{0, \frac{p^2}{1-p}\} \\ > 0 & \text{if } 0 < \Delta < \frac{p^2}{1-p} \\ < 0 & \text{otherwise} \end{cases} .$$

Proof We set

$$h(\Delta) := P^{O,\overline{R}}(z_1, z_2|c^{\pm\Delta}) = \frac{(p+\Delta)(p-\Delta)}{1-p-\Delta} = \frac{p^2-\Delta^2}{1-p-\Delta}$$

and observe that

$$\begin{aligned} P^{O,\overline{R}}(z_1, z_2|c^{\pm\Delta}) - P^{O,\overline{R}}(z_1, z_2|c^{\pm 0}) &= h(\Delta) - h(0) \\ &= \frac{(1-p)(p^2-\Delta^2) - (1-p-\Delta)p^2}{(1-p-\Delta)(1-p)} \\ &= \frac{\Delta(p^2-(1-p)\Delta)}{(1-p-\Delta)(1-p)} \ . \end{aligned}$$

The denominator of the latter expression is strictly positive. Moreover

$$\Delta(p^2-(1-p)\Delta) \begin{cases} = 0 & \text{if } \Delta \in \{0, \frac{p^2}{1-p}\} \\ > 0 & \text{if } 0 < \Delta < \frac{p^2}{1-p} \\ < 0 & \text{otherwise} \end{cases} ,$$

which accomplishes the proof of Fact 2. ■

Fact 3: Suppose that $0 \le \Delta < p < 1/2$. Let $c^{\pm\Delta}$ be the concept whose parameters are given by (10). Then

$$P^{O,R}(z_1, z_1, z_2|c^{\pm\Delta}) - P^{O,R}(z_1, z_1, z_2|c^{\pm 0}) \begin{cases} = 0 & \text{if } \Delta \in \{0, \frac{1}{2}\sqrt{5}-1)p\} \\ > 0 & \text{if } 0 < \Delta < \frac{1}{2}(\sqrt{5}-1)p \\ < 0 & \text{otherwise} \end{cases} .$$

Proof Let $0 < \delta < 1$ be given by $\Delta = \delta p$ and note that

$$P^{O,R}(z_1, z_1, z_2|c^{\pm\delta p}) = (p+\delta p)^2 \cdot (p-\delta p) = (1+\delta)^2 \cdot (1-\delta) \cdot p^3 = (1+\delta-\delta^2-\delta^3) \cdot p^3 \ .$$

It follows that

$$P^{O,R}(z_1, z_1, z_2|c^{\pm\delta p}) - P^{O,R}(z_1, z_1, z_2|c^{\pm 0}) = \delta \cdot (1-\delta-\delta^2) \cdot p^3 \ .$$

Furthermore

$$\delta \cdot (1-\delta-\delta^2) \begin{cases} = 0 & \text{if } \delta \in \left\{0, \frac{1}{2}(\sqrt{5}-1)\right\} \\ > 0 & \text{if } 0 < \delta < \frac{1}{2}(\sqrt{5}-1) \\ < 0 & \text{otherwise} \end{cases} .$$

We may conclude from this discussion that (12) is valid. ■

Fact 4: Suppose that $0 < p < 1/2$ and $1 \le t < \frac{1-p}{p}$. Let $c^{(t)}$ be the concept whose parameters are given by (13). Then $P^{\overline{O},\overline{R}}(z_1, z_2|c^{(t)})$ is strictly increasing with t.

Proof Set

$$\begin{aligned} h(t) &:= \frac{P^{\overline{O},\overline{R}}(z_1, z_2|c^{(t)})}{p^2} = \frac{1}{1-pt} + \frac{1}{1-p/t} = \frac{1}{1-pt} + \frac{t}{t-p} \\ &= \frac{(t-p)+(1-pt)t}{(1-pt)(t-p)} = \frac{2t-pt^2-p}{(p^2+1)t-pt^2-p} \ . \end{aligned}$$

It suffices to show that $h(t)$ is strictly increasing with t. To this end, we compute the first derivative:

$$h'(t) = \frac{(2-2pt)\cdot((p^2+1)t-pt^2-p)-(2t-pt^2-p)(p^2+1-2pt)}{(1-pt)^2\cdot(t-p)^2} \ .$$

The denominator is strictly positive. After an application of the distributive law and some cancellation, the numerator has the form

$$f(t) := p(1-p^2)(t^2-1) \ .$$

Hence the numerator equals 0 for $t = 1$ and is strictly positive for $t > 1$. It follows that $h(t)$ with $t \ge 1$ is strictly increasing. ■

References

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

Baxter S. Eaves, Jr. and Patrick Shafto. Toward a general, scaleable framework for Bayesian teaching with applications to topic models. *CoRR*, 2016. URL `http://arxiv.org/abs/1605.07999`.

Shaun Fallat, David Kirkpatrick, Hans U. Simon, Abolghasem Soltani, and Sandra Zilles. On batch teaching without collusion. *Journal of Machine Learning Research*, 23:1–32, 2022.

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*, pages 2973–2979, 2022. URL `https://doi.org/10.24963/ijcai.2022/412`.

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

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

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

Brigt Arve Toppe Håvardstun, Cèsar Ferri, José Hernandez-Orallo, Pekka Parviainen, and Jan Arne Telle. XAI with machine teaching when humans are (not) informed about the irrelevant features. In *ECML*, 2023. To appear.

David G. Kirkpatrick, Hans U. Simon, and Sandra Zilles. Optimal collusion-free teaching. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of Machine Learning Research (ALT 2019)*, volume 98, pages 1–23, 2019. URL `http://proceedings.mlr.press/v98/kirkpatrick19a/kirkpatrick19a.pdf`.

Farnam Mansouri, Hans Simon, Adish Singla, and Sandra Zilles. On batch teaching with sample complexity bounded by vcd. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15732–15742. Curran Associates, Inc., 2022.

Stephen H. Muggleton. Learning from positive data. In Stephen H. Muggleton, editor, *Inductive Logic Programming, 6th International Workshop, ILP 1996*, volume 1314 of *Lecture Notes in Computer Science*, pages 358–376. Springer, 1996. URL `https://link.springer.com/content/pdf/10.1007/3-540-63494-0_65.pdf`.

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. URL `http://www.sciencedirect.com/science/article/pii/S0010028514000024`.

Ayumi Shinohara and Satoru Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–348, 1991.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & its Applications*, 16 (2):264–280, 1971.

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. URL `https://www.scottchenghsinyang.com/paper/YangShafto_NIPS_2017.pdf`.

Jerry Zhu. Machine teaching for bayesian learners in the exponential family. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An overview of machine teaching. *CoRR*, 2018. URL `https://doi.org/10.48550/arXiv.1801.05927`.

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