MAP-and MLE-Based Teaching[⋆]

Hans Ulrich Simon[1,2][0000−0002−1587−0944]
and Jan Arne Telle[3][0000−0002−9429−5377]

[1] Max-Planck Institute for Informatics, Saarbrücken, Germany
[2] Department of Mathematics, Ruhr-University Bochum, Germany
hsimon@mpi-inf.mpg.de
https://www.ruhr-uni-bochum.de/lmi/simon/
[3] Department of Informatics, University of Bergen, Norway
jan.arne.telle@uib.no
https://www.uib.no/en/persons/Jan.Arne.Telle

Abstract. Imagine a learner L who tries to infer a hidden concept from a collection of observations. Building on the work [4] of Ferri et al., 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. 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. Third, we characterize the MAP-teaching dimensions associated with optimally parameterized MAP-learners graph-theoretically. As a by-product of this characterization, these dimensions can be bounded from above by the so-called antichain number of C, the VC-dimension of C and related combinatorial parameters. The third result is shown only for the (important!) special case where concepts are subsets of a domain and observations are 0,1-labeled examples.

Keywords: machine teaching · MAP · MLE · bipartite matchings

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

[⋆] Supported by the Norwegian Research Council, project Machine Teaching for XAI.

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 [8], trustwhorthy AI [15] and pedagogy [12].

Various models of machine teaching have been proposed, e.g. the classical teaching model [6], the optimal teacher model [1], recursive teaching [16], preference-based teaching [5], or no-clash teaching [9, 3]. 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 [11] where examples are sampled based on likelihoods for a target concept, to Shafto et al [12] who calls this pedagogical sampling and leads into Bayesian Teaching [2, 13], to the Bayesian learners of Zhu [14] 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 an extension of the model that was introduced in [4]. As already observed in [4], the condition for collusion-avoidance from [7] 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. As the authors of [4], 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.

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 MAP-$\mathrm{TD}_L^{\alpha,\beta}(C)$ where $(\alpha, \beta) \in \{O, \overline{O}\} \times \{R, \overline{R}\}$ indicates the underlying sampling mode. Then MAP-$\mathrm{TD}^{\alpha,\beta}(C) = \min_L$ MAP-$\mathrm{TD}_L^{\alpha,\beta}(C)$ 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 3.1 for details.
2. The sampling modes (O, R) and $(\overline{O}, R)$ are equivalent, which implies that MAP-$\mathrm{TD}^{O,R}(C) =$ MAP-$\mathrm{TD}^{\overline{O},R}(C)$ for each concept class C. Furthermore, 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 MAP-$\mathrm{TD}_L(C)$ 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.

3. For a (properly defined) bipartite graph $G(C)^{\alpha,\beta}$ associated with C and $(\alpha,\beta) \neq (O,R)$, one gets[4]

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

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.[5] Fig. 1 visualizes this result.

The third result holds (in this strength) only for the special case where C is a family of subsets of a domain X and $Z = X \times \{0,1\}$ is the set of $0,1$-labeled examples (but is partially true in the general case). As mentioned in the abstract already, the SMN-characterization of MAP-TD implies that MAP-TD can be upper-bounded by the antichain number of C, the VC-dimension of C and related parameters.

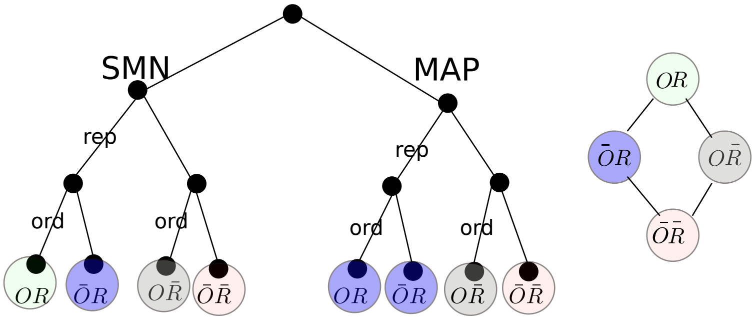

Fig. 1. For any binary concept class $C \subseteq 2^X$ and $0,1$-labeled examples as observations, the tree visualizes the identities in (1). Using the same color for the two leftmost leaves in the MAP-subtree is justified by the equivalence of the modes (O,R) and $(\overline{O},R)$. A parameter represented 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. 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 that parameters represented in different colors can generally have different values.

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

[5] Some bounds on MLE-TD numbers in terms of SMN numbers are already found in [4], but no results that hold with equality (as in (1)) are proven there.

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. 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. 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.

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. multi-set) 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 $Z_c = \{z \in Z : c \models z\}$. The special setting described in the following example is the setting which is used in most papers on machine teaching:

Example 1 (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. We refer to examples with label 1 (resp. with label 0) as *positive* (resp. as *negative*) *examples.* Let the consistency relation be given by

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

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 $c \in C$. Moreover $c \neq c'$ implies that $Z_c \neq Z_{c'}$ so that each concept $c \in C$ is uniquely determined by the full set Z_c of observations that c is consistent with.

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 multi-set $\{z_1, \ldots, z_n\}$.*[6]

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 multi-set. 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:

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

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 $0 \le n \le N^+(Q)$, we define n-fold* (ordered resp. unordered) Q-sampling without replacement *as the following random procedure:*

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 $\frac{q(z)}{1-(q(z_1)+\ldots+q(z_{i-1}))}$.[7]
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:

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

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 multi-sets over Z (including the empty multi-set). $\mathcal{Z}^{O,\overline{R}}$ 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 multi-set or the empty sequence. If A is a finite set or multi-set, then $|A|$ denotes its size where, in case of a multi-set, the multiple occurrences of elements are taken into account. The length of a finite sequence $\mathbf{z}$ is denoted by $|\mathbf{z}|$.

[6] If $n = 0$, then the empty sequence resp. the empty multi-set is returned,

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

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 3 (MAP- and MLE-Learner). *A* MAP-Learner L for C *is given by parameters* $P(z|c) \geq 0$ *and* $P(c) \geq 0$ *for* $z \in Z$ *and* $c \in C$ *such that* $\sum_{c \in C} P(c) = 1$ *and* $\sum_{z \in Z} P(z|c) = 1$*. 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*:*

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

L can be in four different sampling modes. These modes determine the form of L's input and the choice of its output as explained in detail below.

- (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.*[8]
- $(\overline{O},R)$**-mode:** *For every* $n \geq 0$ *and and every multi-set* $A \subseteq Z$ *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 multi-set* $A \in \mathcal{Z}^{\overline{O},R}$*,* 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:** *Set* $N = \min_{c \in C} |Z_c|$*.*[9] *For every* $0 \leq n \leq N$*, 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* $|\mathbf{a}| \leq N$*,* 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* $|\mathbf{a}| > N$*, then also a question mark is returned.*
- $(\overline{O},\overline{R})$**-mode:** *Set again* $N = \min_{c \in C} |Z_c|$*. For every* $0 \leq n \leq N$*, 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* $|A| \leq N$*,* 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* $|A| > N$*, then also a question mark is returned.*

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

[9] Because of the validity condition, $|Z_c|$ equals the number of non-zero parameters in the collection $(P(z|c)_{z \in Z})$.

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 4 (Teacher). *Let $N = \min_{c \in C} |Z_c|$. Suppose that L is a MAP-learner for C that is in sampling mode (α, β). 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 $|I| \le N$.*
2. *$c_0 = \text{arg!max}_{c \in C} \left[P(c) \cdot P^{\alpha,\beta}(I|c)\right]$.*

A couple of observations are in place here (proof omitted):

Remark 1. Suppose that L is a MAP-learner for C which is in sampling mode (α, β). Suppose that T is a teacher for L. Then the following holds for all $c, c' \in C$:

$$L(T(c)) = c\ ,\ P^{\alpha,\beta}(\emptyset|c) = 1\ ,\ P^{\alpha,\beta}(T(c)|c) > 0\ ,\ c \models T(c)\ ,\ (c \ne c' \Rightarrow T(c) \ne T(c'))\ . \tag{2}$$

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

Definition 5 (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. If, in this definition, the MAP-learners are replaced by MLE-learners, we obtain the corresponding notions for MLE-based learning and teaching. Specifically, the* MLE-teaching dimension of C given L and (α, β) *is denoted as* $\text{MLE-TD}_L^{\alpha,\beta}(C)$ *and the* MLE-teaching dimension of C with respect to sampling mode (α, β) *is denoted as* $\text{MLE-TD}^{\alpha,\beta}(C)$.

3 Basic Results on the MAP-Based Teaching Model

We first discuss some natural monotonicity properties and, afterwards, we show the pairwise incomparability of three of the sampling modes (and note the equivalence of (O, R)- and the $(\overline{O}, R)$-mode).

3.1 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 sketch their proofs. All results in this section are formulated in terms of MAP-TD. But the corresponding results with MLE-TD at the place of MAP-TD hold es 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 $c \models' z$. Here is the main result of this section:

Theorem 1. *1. If $(C', Z', \models')$ is an extension of $(C, Z, \models)$ with $Z = Z'$, then*

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

2. If $(C', Z', \models')$ is an extension of $(C, Z, \models)$ with $C = C'$, then

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

Proof. 1. It is sufficient to show that each MAP-learner L' for $(C', Z, \models')$ can be transformed into a MAP-learner L for $(C, Z, \models)$ such that

$$\text{MAP-TD}_L^{\alpha,\beta}(C, Z, \models) \le \text{MAP-TD}_{L'}^{\alpha,\beta}(C', Z, \models') \enspace . \tag{3}$$

Suppose that L' is given by parameters $(P(c))_{c \in C'}$ and $(P(z|c))_{z \in Z, c \in C'}$. Then define L as the learner for $(C, Z, \models)$ with parameters $(P(c))_{c \in C}$ and $(P(z|c))_{z \in Z, c \in C}$. It is obvious that L satisfies (3).

2. It is sufficient to show that each MAP-learner L for $(C, Z, \models)$ can be transformed into a MAP-learner L' for $(C, Z', \models)$ such that

$$\text{MAP-TD}_{L'}^{\alpha,\beta}(C, Z', \models) \le \text{MAP-TD}_L^{\alpha,\beta}(C, Z, \models') \enspace . \tag{4}$$

Suppose that L is given by parameters $(P(c))_{c \in C}$ and $(P(z|c))_{z \in Z, c \in C}$. Then define L' as the learner for $(C, Z, \models)$ with parameters $(P(c))_{c \in C}$ and

$$P'(z|c) = \begin{cases} (1-\varepsilon) \cdot P(z|c) & \text{if } z \in Z \\ \frac{\varepsilon}{|Z'_c \setminus Z|} & \text{if } z \in Z'_c \setminus Z \\ 0 & \text{if } z \in Z' \setminus Z'_c \end{cases} \enspace ,$$

where $\varepsilon \ge 0$, and $\varepsilon = 0$ iff $Z'_c \subseteq Z$. It follows from a simple continuity argument that L' satisfies (4) provided that $\varepsilon > 0$ is sufficiently small.[10] □

3.2 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 $\text{MAP-TD}_L^{\alpha,\beta}(C) \le \text{MAP-TD}_L^{\alpha',\beta'}(C)$. We say they are *equivalent* if they mutually dominate each other. We say they are *incomparable* if none of them dominates the other one. We start with an easy observation (proof omitted):

Remark 2. The sampling modes (O, R) and $(\overline{O}, R)$ are equivalent.

[10] In case of $Z'_c \not\subseteq Z$, we cannot set $\varepsilon = 0$ because the parameters have to satisfy the validity-condition.

The equivalence of these modes implies that

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

Theorem 2. *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

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

Suppose that $0 < |\Delta| < p < 1/2$. The following facts are easy to verify:

1. $P^{O,R}(z_1, z_2)|c^{\pm\Delta})$ and $P^{\overline{O},\overline{R}}(z_1, z_2|c^{\pm\Delta})$ are both strictly decreasing when $|\Delta|$ is increased, which implies that $\Delta = 0$ is the unique maximizer. Loosely speaking, in sampling modes (O,R) and $(\overline{O},\overline{R})$, there is an incentive to split the total mass $2p$ as evenly as possible among z_1 and z_2. We will refer to this as the "even-split argument".
2. In sampling mode $(O,\overline{R})$, however, there is an incentive to split the total probability mass $2p$ not evenly among z_1 and z_2 but slightly in favor of z_1. More precisely, the following holds:

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

Note furthermore 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 as the "monotonicity argument". Theorem 2 is a direct consequence of the following three lemmas.

Lemma 1. *Let L be the MLE-learner for C with parameters $P(z|c)$ 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 $0 < \Delta_1 < \Delta_2 = \frac{p^2}{1-p} < p < 1/2$. *Then*

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

Lemma 2. *Let L be the MLE-learner for C with 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*

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

Lemma 3. *Let L be the 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 $0 < p < \frac{1}{2}$ *and* $1 < s \le \frac{1-p}{p}$. *Then* $\text{MLE-TD}_L^{\overline{O},\overline{R}}(C) = 2 < \text{MLE-TD}_L^{O,R}(C)$, *provided that* $\varepsilon > 0$ *is sufficiently small.*

We omit the proofs because of space constraints. The proofs of the first two lemmas are easy to accomplish by making use of the monotonicity and the even-split argument.

4 Consistency Graphs and 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

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

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 $\text{SMN}(G(C)^{\alpha,\beta})$ the smallest possible order of a C-saturating matching in $G(C)^{\alpha,\beta}$. Analogously, $\text{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset})$ 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 3. Suppose that $T : C \to \mathcal{Z}^{\alpha,\beta}$ is a mapping which satisfies

$$\forall c,c' \in C : (c \models T(c)) \wedge (c \neq c' \Rightarrow T(c) \neq T(c')) \enspace . \tag{5}$$

Then T is of order at least $\text{SMN}(G(C)^{\alpha,\beta})$. Moreover, if T satisfies (5) and $\emptyset$ is not in the image of T, then T is of order at least $\text{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset})$.

Proof. If T satisfies (5), 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 3. *1. For each sampling mode (α, β), we have*

$$\text{MAP-TD}^{\alpha,\beta}(C) \geq \text{SMN}(G(C)^{\alpha,\beta}) \ \ \textit{and}\ \text{MLE-TD}^{\alpha,\beta}(C) \geq \text{SMN}(G(C)^{\alpha,\beta}_{\neq\emptyset}) \ . \qquad (6)$$

2. If $(\alpha, \beta) = (\overline{O}, R)$, then (6) holds with equality.
3. If $(\alpha, \beta) \neq (O, R)$ and $(C, Z, \models)$ is of the form as described in Example 1, then (6) holds with equality.

Proof. We first verify (6). Let L be a MAP-learner for C and let (α, β) denote its sampling mode. Let T be a teacher for L. An inspection of (2) reveals that T satisfies (5). Moreover, if L is an MLE-learner for C, then $T(c) \neq \emptyset$ for all $c \in C$. Now an application of Remark 3 yields (6).

Because of space constraints, we omit the proof of the second assertion and proceed directly with the third one. Since the second assertion settles the result for sampling mode $(\overline{O}, R)$, it suffices to prove the third assertion for the sampling modes $(\overline{O}, \overline{R})$ and $(O, \overline{R})$. In this short abstract, we sketch only the proof of $\text{MLE-TD}^{O,\overline{R}}(C) \leq \text{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})$. Set $m = |X|$ and let $M : C \rightarrow \mathcal{Z}^{O,\overline{R}} \setminus \{\emptyset\}$ be a C-saturating matching in $G(C)^{O,\overline{R}}_{\neq\emptyset}$ of minimum order. For every $c \in C$, we set $d(c) = |M(c)|$. We fix for each concept $c \in C$ a sequence $\mathbf{z^c} = z^c_1, \ldots, z^c_m$ consisting of all elements of Z_c subject to the constraint that $z^c_1, \ldots, z^c_{d(c)} = M(c)$, i.e., the sequence $\mathbf{z^c}$ 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 define a preliminary learner L with parameters $P(z|c)$. The *interim goal* is that any fixed target concept $c^* \in C$ is a (not necessarily unique) maximizer of $P^{O,\overline{R}}_L(M(c^*)|c)$. In Stage 2, we make some infinitesimal changes of the parameter set resulting in an MLE-learner L_ε with parameters $P_\varepsilon(z|c)$. The *ultimate goal* is to show that each $c^* \in C$ is the unique maximizer of $P^{O,\overline{R}}_{L_\varepsilon}(M(c^*|c))$ provided that $\varepsilon > 0$ is sufficiently small. The MLE-learner L used in Stage 1 is given by the following parameters:

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

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 $Z_c \setminus M(c)$. A nice (easy-to-verify) consequence of the above definition of $P(z|c)$ is that $P^{O,\overline{R}}(M(c)|c) = 2^{-d(c)}$. By a careful analysis (omitted here because of space constraints), we can show that, with this definition of L, the interim goal is achieved. The MLE-learner L_ε used in Stage 2 is given by:

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

The main difference to the old parameter collection is the “extra-bonus” ε that c assigns to the last element $z^c_{d(c)}$ of the sequence $M(c)$. By a careful analysis (omitted here because of space constraints), we can show that, with this definition of L_ε, the ultimate goal is achieved. Note that this implies that we may view M as a teacher for L_ε. It follows that $\text{MLE-TD}^{O,\overline{R}}(C) \le \text{SMN}(G(C)^{O,\overline{R}}_{\neq\emptyset})$. □

For the remainder of the paper, we assume that $(C, Z, \models)$ is of the form as described in Example 1. Note that the third assertion in Theorem 3 implies the correctness of the results which are visualized in Fig. 1. The following corollaries (with proofs based on Theorem 3) provide some supplementary information:

Corollary 1. *1.* $\text{MAP-TD}(C) \le \text{MLE-TD}(C) \le 1+\text{MAP-TD}(C)$. *One of the two inequalities must hold with equality. Both cases can occur.*

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

Proof. We briefly sketch the proof of the second assertion for $(\alpha, \beta) = (\overline{O}, R)$ and $(\alpha', \beta') = (\overline{O}, \overline{R})$.[11] Let $X = \{x_1, \ldots, x_m\}$, let $Z = X \times \{0, 1\}$, let C_m be the powerset of X_m and 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 $|\mathcal{Z}'_2| < |\mathcal{Z}_2|$. From Hall’s theorem, it can be inferred that $G(C_m)^{\overline{O},R}$ admits a $\mathcal{Z}_2$-saturating matching, say M_2. Let C be the set of concepts in C_m having an M_2-partner. By construction: $\text{SMN}(G(C)^{\overline{O},R}) = 2$. For cardinality reasons, namely $|C| = |M_2| = |\mathcal{Z}_2| > |\mathcal{Z}'_2|$, we have $\text{SMN}(G(C)^{\overline{O},\overline{R}}) > 2$. □

The second assertion of this corollary implies that the parameters with different colors in Fig. 1 can generally have different values.

$T : C \to 2^Z$ is called an *antichain mapping for C* if the following holds. First, each concept $c \in C$ is consistent with $T(c)$. Second, the sets $(T(c))_{c\in C}$ form an antichain. The smallest possible order of an antichain mapping for C is called the *antichain number of C* and denoted by $\text{AN}(C)$. It is well known [10] that $\text{AN}(C) \le \text{VCdim}(C)$. It is easy to see that $\text{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset}) \le \text{AN}(C)$. Hence:

Corollary 2. $\text{MLE-TD}^{\overline{O},\overline{R}}(C)$ *is upper-bounded by* $\text{AN}(C)$ *and by* $\text{VCdim}(C)$.

Note that $\text{MLE-TD}^{\overline{O},\overline{R}}(C) = \text{SMN}(G(C)^{\overline{O},\overline{R}}_{\neq\emptyset})$ is the largest among all MAP-, MLE- and SMN-parameters associated with C. Hence any of these parameters is upper-bounded by $\text{AN}(C)$ and $\text{VCdim}(C)$.

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?

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

References

1. Balbach, F.: Measuring teachability using variants of the teaching dimension. Theoretical Computer Science **397**(1–3), 94–113 (2008)
2. Eaves, Jr., B.S., Shafto, P.: Toward a general, scaleable framework for Bayesian teaching with applications to topic models. CoRR (2016), http://arxiv.org/abs/1605.07999
3. Fallat, S., Kirkpatrick, D., Simon, H.U., Soltani, A., Zilles, S.: On batch teaching without collusion. Journal of Machine Learning Research **23**, 1–32 (2022)
4. Ferri, C., Hernández-Orallo, J., Telle, J.A.: Non-cheating teaching revisited: A new probabilistic machine teaching model. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022. pp. 2973–2979 (2022), https://doi.org/10.24963/ijcai.2022/412
5. Gao, Z., Ries, C., Simon, H.U., Zilles, S.: Preference-based teaching. Journal of Machine Learning Research **18**(31), 1–32 (2017)
6. Goldman, S.A., Kearns, M.J.: On the complexity of teaching. Journal of Computer and System Sciences **50**(1), 20–31 (1995)
7. Goldman, S.A., Mathias, H.D.: Teaching a smarter learner. Journal of Computer and System Sciences **52**(2), 255–267 (1996)
8. Håvardstun, B.A.T., Ferri, C., Hernandez-Orallo, J., Parviainen, P., Telle, J.A.: XAI with machine teaching when humans are (not) informed about the irrelevant features. In: ECML (2023), to appear
9. Kirkpatrick, D.G., Simon, H.U., Zilles, S.: Optimal collusion-free teaching. In: Garivier, A., Kale, S. (eds.) Proceedings of Machine Learning Research (ALT 2019). vol. 98, pp. 1–23 (2019), http://proceedings.mlr.press/v98/kirkpatrick19a/kirkpatrick19a.pdf
10. Mansouri, F., Simon, H., Singla, A., Zilles, S.: On batch teaching with sample complexity bounded by vcd. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 15732–15742. Curran Associates, Inc. (2022)
11. Muggleton, S.H.: Learning from positive data. In: Muggleton, S.H. (ed.) Inductive Logic Programming, 6th International Workshop, ILP 1996. Lecture Notes in Computer Science, vol. 1314, pp. 358–376. Springer (1996), https://link.springer.com/content/pdf/10.1007/3-540-63494-0_65.pdf
12. Shafto, P., Goodman, N.D., Griffiths, T.L.: A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology **71**, 55 – 89 (2014), http://www.sciencedirect.com/science/article/pii/S0010028514000024
13. Yang, S.C.H., Shafto, P.: Explainable artificial intelligence via bayesian teaching. In: NIPS 2017 workshop on Teaching Machines, Robots, and Humans. pp. 127–137 (2017), https://www.scottchenghsinyang.com/paper/YangShafto_NIPS_2017.pdf
14. Zhu, J.: Machine teaching for bayesian learners in the exponential family. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013)
15. Zhu, X., Singla, A., Zilles, S., Rafferty, A.N.: An overview of machine teaching. CoRR (2018), https://doi.org/10.48550/arXiv.1801.05927
16. Zilles, S., Lange, S., Holte, R., Zinkevich, M.: Models of cooperative teaching and learning. Journal of Machine Learning Research **12**, 349–384 (2011)